University of Zurich[UZH]

Parallel Corpora for the Investigation of (Variable) Article Use in English
A Construction Grammar Approach

Thesis
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of Doctor of Philosophy

by
Elena Callegaro

Accepted in the fall semester 2017
on the recommendation of the doctoral committee composed of:
Prof. Dr. Marianne Hundt (main supervisor)
Prof. Dr. Martin Volk

Zurich, 2020

Table of contents

List of figures

List of tables

Acknowledgments

Many people have helped throughout these years, and listing all of them would be inconceivable. First and foremost, I would like to express my gratitude to my doctoral supervisors Prof. Dr. Marianne Hundt and Prof. Dr. Martin Volk for giving me the opportunity to be part of this research project, for their suggestions, insightful and critical comments, and for the challenging questions which pushed me to widen my research from various perspectives. I would also like to thank the members and collaborators of the project for their hard work, with a special thanks to Dr. Simon Clematide. Also, this would not have been possible without the generous funding provided by the Swiss National Science Foundation.

My sincere gratitude extends to the kind people that I have met in Zurich, in and outside the English Department. I know I have found new true friends. I also would like to thank all those older friends that have listened to me and cheered me up from different parts of the world during difficult times, in particular my best friends Alessia and Claudia. I am truly grateful to my office colleagues Adina and Rahel, for their help, stimulating discussions, precious advice, and for the funny moments we have shared in the past years. I am immensely grateful to André for his enormous support, endless patience, and for being there every single moment.

Last but certainly not least, I would like to thank my beautiful family for the constant support and for believing in me. Words cannot express how grateful I am. Thank you *mamma*, *papà*, Elisa, *nonna*, and Alvise. *Tutto questo non sarebbe stato possibile senza di voi.*

1 Introduction

1.1 Motivation and purpose

In English, articles are among the most frequently occurring words (Kučera and Francis 1967: 5, Johansson and Hofland 1989: 19). Standard grammars agree on the fact that they are part of the closed category called *determiners* and are considered as the most basic units expressing definiteness and indefiniteness: the definite article generally precedes a noun within a noun phrase (NP) to express definiteness, whereas the indefinite article is used to express indefiniteness. On the other hand, articles are omitted to give a conception of a whole class with a general connotation (Quirk et al. 1985; Biber et al. 1999; Huddleston and Pullum 2002). At first, this differentiation seems clear and straightforward; however, a closer look at article usage promptly reveals how complex articles are. Their use can be influenced by many factors, such as the noun number, the type of noun they precede, or the syntactic function of the NP. Moreover, the presence or absence of an article within a NP can change the meaning of the whole noun phrase (e.g. specific vs. generic, familiar vs. non-familiar, identifiable vs. non-identifiable). Hence, when analysing an NP, article omission is as important as article use. Another significant aspect that makes articles so interesting and intricate is variability without a change in meaning (e.g. *she is (the) President of the company*).

In the literature, articles have been extensively investigated and have received considerable attention due to their complexity in usage. However, to date, much of the research has been mainly descriptive in nature and has primarily focused on the uses of the definite and indefinite article (Berezowski 2009: 1). Furthermore, paucity of research remains with respect to their variable usage (but see Tse 2001, 2003, 2004; Yoo 2007; Hundt 2016, 2018; Callegaro et al. 2019). The main goal of the current study therefore lies in the investigation of variable article use with particular focus on British English (BrE). The study was also encouraged by the fact that zero cases have reportedly become more common in English. Previous research has attested that in English the frequency of bare NPs contexts has increased over time. In their extensive study, Leech et al. (2009) use corpora from the Brown Family[1] to analyse English

[1] The corpora included in the study are the following: the Brown Corpus (American English, 1961), the Lancaster–Oslo/Bergen corpus (British English, 1961), the Freiburg–Brown corpus

language change in the thirty years between the early 1960s and early 1990s. Leech et al. (2009: 209) state that the use of the definite article decreased by 11.1% in American English (AmE) and by 5.4% in BrE. Two particular cases are discussed. The first instance refers to noun phrases postmodified by the preposition *of* (e.g. *the fruit of the coconut palm* and *the behaviour of the patient*). In the corpora, the decline of the preposition *of* is attested in both varieties,[2] and the consequent loss is closely related to the use of two other constructions, namely the *s*-genitive construction (e.g. *the coconut palm's fruit* and *the patient's behaviour*), and the Noun + Noun sequence construction (e.g. *coconut palm fruit* and *patient behaviour*), in which the preposition *of* is omitted together with the definite article. The second instance of article loss refers to appositions, which are typically found in the journalistic genre. Rydén (1975), Bell (1985, 1988) and Jucker (1992) affirm that, over the last century, articles have been increasingly omitted in appositions consisting of two nominals, in contrast to the more common construction in which the apposition postmodifies the noun. To illustrate, the newer construction *midnight bather Brian Best* is now preferred over the more standard construction *Brian Best, the midnight bather* (Leech et al. 2009: 216). In these two cases, therefore, it is noticeable that the NP becomes shorter and more compact. In order to condense the information of an NP, articles are eligible elements for omission. As mentioned before, this phenomenon was the motivation that drove the current investigation, whose aim was thus to find novel bare cases.

The method used to investigate article use/omission is corpus-based. With respect to article use, this methodology is not new. Biber et al. (1999: 266-270) use the Longman Spoken and Written English Corpus (LSWE Corpus) to describe the use of definite noun phrases across various registers (i.e. conversation transcriptions, academic texts, fiction texts, and newspaper texts). However, their analyses only compare the use of the definite and indefinite article and contrast the distribution of the definite article with demonstratives and possessives. Bare NPs are thus not taken into account. Targeting NPs in which an article is omitted is a big challenge; such cases are notoriously hard to obtain with monolingual corpora: the retrieval of bare NPs suffers from low precision. In order to go beyond this difficulty, the current study uses large parallel corpora. Parallel corpora are a powerful and innovative tool

(American English, 1992), and the Freiburg–Lancaster–Oslo/Bergen corpus (British English, 1991).

[2] The preposition *of* declines by 11.3% in AmE and by 4.7% in BrE.

because, with the help of a second language, it is possible to obtain cases in which an article does not occur in the language of interest. Parallel material has been proven to be a valuable resource in various fields (e.g. translation studies, contrastive linguistics, and natural language processing). The corpus used in the current study is called *Europarl – a Parallel Corpus for Statistical Machine Translation* (Koehn 2005, see section 4.1). It consists of the proceedings of the European Parliament and includes both original and translated material in various languages of the European Union.[3] The language pair on which the current analysis focuses is English – German. By using German as a starting point, it is possible to target German chunks in which an article is used and retrieve the aligned equivalent English chunks occurring without an article (see section 4.1.1). German is expected to use articles more frequently than English, as seen in Table 1.1, which raises the chances of finding novel bare cases in English.[4]

	no art. English	**art. English**	***total***
no art. German	120.307	**36.671**	*156.978*
art. German	**67.764**	166.012	*233.776*
total	*188.071*	*202.683*	*390.754*

Table 1.1: Number of NP alignments in *Europarl* with and without an article in English and German (raw numbers).

On a more theoretical level, this study investigates (variable) article use in English from a Construction Grammar (CxG) perspective. Scholars have shown increased interest in this theoretical framework, but there is still a lack of research in relation to English articles. This relatively new theory examines how speakers of a given language use that language and what they know about the language they speak (see Chapter 3). Construction Grammar is used in the current study because it has many advantages. First of all, since the constructional approach is non-compositional, it is able to explain any linguistic unit. Second of all, grammatical constructions can be analysed on different levels of abstractions. Furthermore, with respect to the use of linguistic corpora, both sub-regularities and more general patterns are equally taken into account and can be analysed in depth. Finally, on a more general note,

[3] Note that the corpus does not allow for diachronic studies; therefore, research is based on synchronic investigation.
[4] The data included in Table 1.1 were available only at a later stage of the project.

Construction Grammar allows for a more detailed and analytic look at (variable) article use than standard grammars take and thus distances itself from broader, descriptive approaches.

The current study therefore tries to investigate whether it is possible to build up generalizations with a bottom-up approach. In other words, whether it is possible to study article use starting directly from the construct level (i.e. occurrences retrieved from the corpus) and moving towards more abstract levels. Furthermore, the current study tries to prove the usefulness of large parallel corpora for the investigation of language variation. In the context of the overarching project, the goal of the data analysis is twofold: on the one hand, to explore the contrastive opportunities that the parallel corpus offers, and on the other hand, to extract relevant data for a construction grammar study of (variable) article use in English. Also, it aims to validate the application of Construction Grammar to a corpus approach. Overall, the outcomes of the analysis will show that all these questions have positive responses. Using German translations as the starting point, the retrieval process will produce a sample containing original English instances with bare NPs that mainly contain abstract nouns. As described by Rowlinson (1994: 87-90), abstract nouns are generally preceded by an article in German, while they tend to occur as bare NPs in English (see section 2.6). However, the results will reveal that this is not always the case. With the constructional theory, it will be possible to interpret both the general tendencies and the cases with lower frequency. Moreover, it will shed new light on English article use from a construction perspective. This study thus proves that the combination of parallel corpora with data-driven research and statistical modelling of variation, and the application of the constructional framework yields meaningful results. Finally, it plays an important role in addressing the issue of variable article use and provides new insights into different research areas, i.e. language variation, contrastive linguistics, translation studies, corpus linguistics, computational linguistics, and Construction Grammar.

From this motivation, the following research questions and hypotheses emerge:

1. To what extend are parallel corpora useful for the investigation of article variation?

   It is hypothesized that parallel corpora will greatly facilitate the retrieval of bare NPs and therefore the discussion of variable article use in British English.

2. More specifically, since the Europarl corpus contains transcriptions of parliamentary speeches, can it be considered suitable for linguistic research?
   It is hypothesized that Europarl, due to its contextually constrained nature and assumed careful transcriptions (based on its legal quality), provides useable material for linguistically motivated research.
3. Does retrieving English bare NPs via aligned German non-bare NPs yield novel bare cases in British English?
   It is hypothesized that this methodology will yield English bare noun phrases previously unaddressed in article variability research.
4. From the Construction Grammar perspective, is it possible to build up generalizations with a bottom-up approach based on corpus data (i.e. study article use starting from the construct level and moving towards more abstract levels)?
   It is hypothesized that by focussing on corpus evidence at the construct level of CxG previous constructional models can be improved by building the model in a bottom-up fashion.

1.2 Outline of chapters

The overall structure of the study takes the form of two main parts: the first focuses on the background information, while the second provides the analysis of the data. The first part begins with Chapter 2, which gives a review of literature with respect to the articles. The main aspects of the definite article, indefinite article, and article omission in both English and German are presented. It then continues with a short summary of the most relevant theories on definiteness and the differences on article use between English and German. The last two sections of Chapter 2 focus on the review of the main studies on variable article use in English and the differences between BrE and Irish English (IrE). The differentiation between these two varieties is important because the proceedings collected in *Europarl* include both British and Irish speakers, and literature has shown that article use in IrE considerably differs from BrE (see e.g. Hickey 2007, Siemund 2013).

Chapter 3 gives an overview of the Construction Grammar framework and presents its main concepts. It then looks at how the constructional approach has been used to describe English article use and focuses on the limitations of the suggested

theories. Finally, it examines how Construction Grammar has been applied in the field of language variation.

Chapter 4 is concerned with the corpus data and the methodology used for this study. It begins with a focus on the contribution of (parallel) corpora to linguistic investigation and continues with the discussion on how these can be useful for the investigation of (variable) article use in English. It moves on to describing the *Europarl* corpus, its newer and improved version called *CoStEP*, and how this was aligned and annotated. This is followed by two preliminary case studies: the former analyses the reliability of the corpus by comparing the transcriptions with the original video recordings, the latter compares original texts and translations and examines the differences regarding article use.

The second part of the dissertation begins with Chapter 5, which presents the first case study conducted during the current project. The main aim of this chapter is to explore the opportunities a parallel corpus can offer to linguistic analysis. It investigates article use using a lexeme-based approach, namely with four collective nouns (i.e. *Parliament*, *Council*, *Committee*, and *people*) and compares the results between English and German. In addition, it investigates the differences between BrE and IrE.

By contrast, Chapter 6 explores English articles using a data-driven approach and explores the nature of the retrieved bare NPs in English. This is followed by the findings of the research, and the follow-up analysis in which abstract nouns are further investigated in comparison to a second construction, i.e. the *of*-CONSTRUCTION. The chapter then continues with an analysis of (variable) article use from a constructional point of view, strictly based on empirical evidence (i.e. using a bottom-up method).

Finally, Chapter 7 provides the summary of the results, a critical discussion of the positive and negative aspects of the approaches used in the current study (i.e. the limitations of the corpus), and a brief account of further research.

Part I – Background

2 Review of literature

2.1 The articles

In many languages, articles are usually the most frequent words. Articles occur in various languages, including English and German, while other languages, such as Russian or Finnish, have no articles. In the sample consulted by Dryer (2013), 198 out of 620 languages have neither definite nor indefinite articles, and 45 languages have an indefinite article but no definite article. Finally, 98 out of 534 languages have a definite but no indefinite article. Therefore, English belongs to the majority of languages that have an article. However, this is not true for all of its history, because the definite article derives from the demonstrative pronoun, while the indefinite article comes from the numeral *one* (Burrow and Turville-Petre 2005: 26-27; Baker 2011: 44).

English standard grammars agree on the fact that articles are part of a limited class called *determiners*. These are words whose function is to specify the references and applications of a noun. Both Quirk et al. (1985: 253) and Biber et al. (1999: 258) distinguish between *predeterminers*, *central determiners*, and *postdeterminers*. They place articles, together with demonstratives and possessives, in the second subgroup, i.e. central determiners. According to *Collins Online* (2017) articles lack "independent meaning but may serve to indicate the specificity of reference of the noun phrase with which [they] occur." Huddleston and Pullum (2002: 368) also state that articles "provide the most basic expression of definiteness and indefiniteness."

The articles *the* in English and *der*, *die*, *das* in German are generally used as part of a noun phrase and express definiteness. In contrast, *a* or *an* in English and *ein*, *eine* in German generally express indefiniteness (or non-definiteness). Despite this categorisation, Crystal (2008: 241) states that, due to possible linguistic and extra-linguistic contextual variables, the distinction between definite and indefinite can prove problematic, because it is not always straightforward when to use one and when to use the other one. Another important characteristic concerning articles is that they can be used with *specific* or *generic* reference: a specific noun phrase denotes a

distinct entity of a class, while a generic noun phrase "refers to a whole class rather than to an individual person or thing" (Biber et al. 1999: 265). For both English and German, these dimensions, i.e. definiteness vs. non-definiteness (or identifiability vs. non-identifiability) and specificity vs. non-specificity (or genericness) are interrelated and expressed through the presence or absence of articles. This relationship is shown in Table 2.1, which is an adaptation of Hentschel's (2010: 37) classification. The definite article connects the dimensions specificity and definiteness, while article omission merges genericness with definiteness. On the other hand, the indefinite article has two possible combinations, namely non-definiteness and specificity or non-definiteness and genericness.

	specific	**non-specific**
identifiable	**definite article** *Der Mond scheint.* *The moon shines.*	**article omission** *Auf der Strasse liegt Schnee.* *There is snow on the street.*
unidentifiable	**indefinite article** *Ich habe mir einen Krimi gekauft.* *I bought myself a crime novel.*	**indefinite article** *Kauf dir doch einen Krimi!* *Buy yourself a crime novel!*

Table 2.1: Relation between specificity, non-specificity and identifiability, non-identifiability in English and German (Hentschel 2010: 37, adapted).

Thus, at a first glance, articles in English and German appear to be a simple and uncomplicated (parallel) binary system between definite and indefinite, but a closer look in the following sections will reveal that "the use of articles presents a great many intricate problems" (Jespersen 1949: 404). This chapter provides a summary of previous descriptions of articles in English and German grammars, the main theories that have been developed to explain definiteness, previous research on (variable) article use, and the difference of article use between BrE and IrE.

2.2 The definite article

As pointed out by Biber et al. (1999: 263), the definite article "combines with both countable and uncountable nouns. It specifies that the referent of the noun phrase is assumed to be known to the speaker and addressee." Quirk et al. (1985: 265) also affirm that it "refer[s] to something which can be identified uniquely in the contextual or general knowledge shared by the speaker and hearer." The definite article marks

the noun phrase as unique, specific, and clearly defined. Quirk et al. (1985: 266-272) list several situations in which the definite article is used, which can be applied to German, too[5]: the *immediate situation* refers to a situation in which the speaker and the hearer share the same knowledge, as in (1); in a *larger situation* the referent is part of general knowledge (e.g. *the Prime Minister, the Pope, die Sonne, der Himmel*); in *direct anaphoric reference*, the information is previously mentioned, as in (2); with *indirect anaphoric reference*, the hearer is given the information indirectly, as in (3); in *cataphoric reference*, the relevant information follows the head noun, as in (4); *sporadic reference* is concerned with articles that refer to an institution of human society, as in (5); in the *logical situation* the use of *the* is given by a logical interpretation, as in (6); and finally, *the* is used with reference to body parts, as in (7).

(1) *The roses* are very beautiful. (said in the garden)
(*Die Rosen* sind sehr hübsch.)
(2) John bought a TV and *a video recorder*, but he returned *the video recorder*.
(John hat einen Fernseher und *einen Videorekorder* gekauft, aber er hat *den Videorekorder* zurückgebracht.)
(3) John bought *a bicycle*, but when he rode it one of *the wheels* came off.
(John hat *ein Fahrrad* gekauft, aber als er es gefahren ist, ist eines *der Räder* abgefallen.)
(4) *The President of Mexico* is to visit China.
(*Der Präsident von Mexico* wird China besuchen.)
(5) My sister goes to *the theatre* every month.
(Meine Schwester geht jeden Monat *ins Theater*.)
(6) When is *the first flight* to Chicago tomorrow?
(Wann geht morgen *der erste Flug* nach Chicago?)
(7) Mary banged herself on *the forehead*.
(Mary hat sich *die Stirn* angeschlagen.)

As indicated previously, the definite article can also carry generic reference, as shown in (8) and (9):

(8) No one knows precisely when *the wheel* was invented.
(Niemand weiß genau wann *das Rad* erfunden wurde.)
(9) *The Welsh* are fond of singing.
(*Die Waliser* singen gerne.)

[5] The English examples of sections 2.2, 2.3 and 2.4 are taken from Quirk et al. (1985).

The meaning of the definite article has been described in a large body of literature and, over the years, many scholars have tried to refine what definiteness is. For instance, Table 2.1 above shows that the two main aspects of the definite article are specificity and identifiability, whereas Birner and Ward (1994) talk about familiarity and uniqueness. An overview of the main theories developed on this topic is given in section 2.5.

2.3 The indefinite article

The indefinite article – *a/an* in English and *ein/eine* in German – is used with singular countable nouns. Jespersen (1949: 419) states that the English indefinite article "denotes one member of a class or species concerned, but it does not indicate which member." Therefore, it is not as specific as the definite article and is normally used to introduce a new entity in the discourse between the speaker and the hearer (Biber et al. 1999: 260). Quirk et al. (1985: 272) define it as

> the 'unmarked' article in the sense that it is used […] where the conditions for the use of *the* do not obtain. That is, *a/an X* will be used where the reference of X is not uniquely identifiable in the shared knowledge of speaker and hearer. Hence *a/an* is typically used when the referent has not been mentioned before, and is assumed to be unfamiliar to the speaker and hearer.

Example (10) shows the use of the indefinite article with a specific reference. On the other hand, example (11) shows the generic use of the article, in which the referent is the representative member of an entire class.

(10) *An intruder* has stolen *a vase*. *The intruder* stole *the vase* from *a locked case*. *The case* was smashed open.
(*Ein Einbrecher* hat *eine Vase* gestohlen. *Der Einbrecher* hat *die Vase* aus *einer verschlossenen Vitrine* gestohlen. *Die Vitrine* wurde zertrümmert.)

(11) The best way to learn *a language* is to live among its speakers.
(Die beste Art *eine Sprache* zu lernen ist unter ihren Sprechern zu leben.)

From this, one can conclude that the functions of the indefinite article in English and German are relatively undisputed.

So far this chapter has focused on the presence of an article within an NP. The following section will analyse in which contexts an article is omitted and what functions article omission has.

2.4 Omission of the article

At times, nouns occur without articles (i.e. bare NPs). Various scholars have addressed the fact that this does not simply constitute the absence of an article but can be considered more complex under the surface (e.g. Berezowski 2009). Some talk about the *zero article* and separate it from the *null article*. The former "precedes mass nouns and plural count nouns", while the latter "precedes singular proper nouns and some singular count nouns" (Yoo 2009: 269). For instance, Chesterman (1991) uses this terminology when arranging the English articles on a scale of definiteness. More specifically, he locates *a* and *the*[6] in the intermediate positions, and defines the zero article as the most indefinite (e.g. *olives*, *cheese*), and the null article as the most definite article (e.g. *John*, *Helsinki*). Furthermore, Chesterman (1991: 182) describes both zero and null as unmarked and states that "it is pragmatically unnecessary to mark forms which are already 'conceptually clear' in some relevant sense." Curme (1970: 62), on the other hand, makes no distinction between null and zero articles and simply notes that "[t]he absence of the article suggests something indefinite or the general conception of class or kind with only a general characterization." With specific reference, articles are generally omitted with proper nouns (e.g. *Paris, Moritz*), and nouns that refer to a unique role or task. However, in both English and German, these can alternate with the definite article, as in (12).[7]

(12) Maureen is *(the) captain* of the team.
(Maureen ist *(die) Spielführerin*.)

Quirk et al. (1985: 276-281) ascribe the situations in which bare NPs occur in English to idiomatic usage: neither a definite nor indefinite article is used with some institutions (e.g. *be in town, go to school*), means of transportation and communication (e.g. *travel by car*), times of the day and night (e.g. *at dawn*), seasons (e.g. *in (the) spring*), meals (e.g. *after lunch*), illnesses (e.g. *anaemia, diabetes*)[8], parallel structures (e.g. *hand in hand*), and fixed phrases involving prepositions (e.g. *at home*).

[6] Note that Chesterman (1991) collocates *some* in the intermediate position together with the indefinite and definite articles.
[7] In German, an alternation with the indefinite article can be found with nouns that refer, for instance, to nationality or regional provenance, due to regional differences, as in *Er ist (ein) Engländer* or *Sie ist (eine) Heidelbergerin* (Dudenredaktion 2005: 339).
[8] Note, however, that the expression *I've got a nasty cold* is possible (McIntosh 2002: 17).

For German, Curme (1970: 67-68) describes the three main circumstances where an article is generally omitted as follows:

> (1) sometimes when the noun contains an abstract idea and […] the general conception of a class or kind and hence does not designate a definite object; (2) when […] the object is already sufficiently defined […], and (3) in many set expressions and proverbs coined in an early period when the article was little used.

Additionally, in both English and German, NPs are bare with plural count nouns, as in (13), and non-count nouns, as in (14); in these cases the absence of the article implies a generic reference.

(13) *Cigarettes* are bad for your health.
(*Zigaretten* sind schlecht für deine Gesundheit.)
(14) *Hunger* and *violence* will continue to mark the future of mankind / humanity.
(*Hunger* und *Gewalt* werden weiterhin die Zukunft der Menschheit prägen.)[9]

A final observation regards the distinction of two functions of bare NPs in English. Christophersen (1939: 36) and Jespersen (1949: 438-439) distinguish between two types of usage: *parti-generic* and *toto-generic*.[10] Lyons (1991: 321) and Siepmann (2001: 1) explain that the former indicates an indefinite quantity or number (e.g. *we had tea*), while the latter designates the class as a whole, rather than all its members, and might therefore be considered as *generic* (e.g. *lead is heavier than iron*).

2.5 On definiteness

In English, definiteness and indefiniteness seem to be clearly marked. More specifically, the definite article, together with demonstratives (e.g. *this, that*), possessives (e.g. *my, your, her*), personal pronouns (e.g. *I, we, they*), proper nouns (e.g. *Paul, Italy*), and some quantifiers (e.g. *all, every*), are used for/as definite NPs; while the indefinite article, article omission and other quantifiers (e.g. numerals such as *some, any, one*) are used for indefinite NPs (Prince 1992: 299). Hence, as Prince

[9] For the difference in article use in *mankind/humanity* and *Menschheit* see section 2.6.
[10] Christophersen (1939: 36) adds a third type called the *nulli-generic*, in which the zero article is used in negative sentences (e.g. *I have not tasted food for three days* or *they never get rain in summer*). However, Jespersen (1949: 440-441) considers this third type not necessary, and believes that in most cases the *nulli-generic* rather refers to the *parti-generic*.

(1992: 299) states, "whether a given NP is formally definite or indefinite is decidable, entirely and exclusively, on the basis of the form of that NP." However, it is possible to use a definite form and have an indefinite meaning and vice versa (Prince 1992: 300); for instance, as already shown in (8) and (9), the definite article is used in an NP with generic reference. The discussion of article variability in Chapter 6 is primarily based in the classification introduced by Chesterman (1991) that was mentioned in the previous section, while the following chapters more concerned with articles' surface forms. What follows is a brief theoretical addition for the sake of completeness that will only be marginally referred to throughout the rest of the work presented here.

Scholars have analysed definiteness from different perspectives and there is no agreed definition on what it constitutes. There are, in fact, two main approaches to definiteness, namely the *familiarity theory* and the *uniqueness approach* (also known as *uniqueness identifiability approach* or *quantifier theory*). The familiarity approach was initially introduced by Christopherson (1939) and was then embraced by other scholars (e.g Strawson 1950; Bolinger 1977; Heim 1983; Elbourne 2010). This theory "is based on the idea that the referent is known to the addressee" (Gisborne 2012: 7). In other words, the "felicitous use of *the* requires only that the referent have [sic!] been introduced into the discourse" (Epstein 2001: 336). On the other hand, with the uniqueness approach, first adopted by Russell (1905) and later defended, for instance, by Neale (1990) and Gisborne (2012), the referent needs to be identifiable to the hearer. As explained by Epstein (2001: 336), "[f]or a referent to be identifiable, it is generally agreed that the referent must be unique, i.e., the only entity of that type within the discourse model." These theories are closely related, because they share the same goal, i.e. the attempt to explain the function of the definite article. As Birner and Ward (1994: 96) affirm:

> there is a great deal of overlap between the set of entities that are (presumed to be) familiar to a hearer and the set of entities that are (presumed to be) uniquely identifiable to the hearer, since an entity typically must be familiar in a given discourse in order to be identifiable.

However, even though familiarity and uniqueness are connected to each other, they are not equivalent. The definite article can also be used to refer to an entity that was not previously mentioned by the speaker. For this reason, Birner and Ward (1994)

believe that familiarity is not necessary, as shown in (15)[11], where the referent is unique but not familiar. On the other hand, familiarity is not a sufficient condition either for the correct use of the definite article. For instance, in (16) the hearer does not know which of the two grants the members of the department are referring to.

(15) If you're going to the bedroom, would you mind bringing back *the big bag potato chips that I left on the bed*?
(16) Professor Smith and Jones are rivals in the English Department, and each of them has received a major research grant for next year. #The other members of the department are very excited about *the grant*.

Similarly, uniqueness is not necessary for the appropriate use of the definite article, as shown in (17), where the referent is familiar but not unique. But, contrary to familiarity, uniqueness is considered sufficient for the correct use of the definite article, as in (18), where the speaker is sure that the entity being referred to is identifiable by the hearer.

(17) [Hotel concierge to guest, in a lobby with four elevators] You're in Room 611. Take *the elevator* to the sixth floor and turn left.
(18) *The King* is dead. Long live *the king*!

Birner and Ward (1994: 101) conclude that

> no single factor proposed – familiarity, uniqueness […] – can alone account for the full distribution of the definite article in English. In particular, pragmatic factors such as the inferred intent of the speaker and the differentiability of referents in context contribute crucially to the interpretation of the definite article.

According to Epstein (2001: 333), neither of these theories can "provide necessary and sufficient conditions for the use of the definite article in English." He proposes a new and more dynamic theoretical account to definiteness, namely the *mental space approach*. Epstein (2001: 348-363) claims that familiarity and identifiability are not the only functions of the definite article. On the contrary, it also indicates: a) the discourse prominence of an entity (i.e. the referent becomes the main focus of attention in an episode and will therefore be the main topic in the subsequent discourse); b) the status of an entity as a role function (i.e. when the definite article is used as a role, it allows the speaker to achieve a particular goal in specific contexts);

[11] Examples (15), (16), (17), and (18) are taken from Birnen and Ward (1994: 93-95).

and c) the non-canonical point of view (i.e. the shift of point of view of a third person that can be either a fictional narrator or a discourse protagonist). In short, therefore, Epstein (2001: 334) suggests that the uses of the definite article "mark the 'accessibility' of a discourse referent – more specifically, a low degree of accessibility." Furthermore, he claims that the definite article itself "is a grammatical signal contributing to both the construction and retrieval of mental entities." In other words, in order to identify an NP, the speaker induces the hearer(s) to access the referent via an access path, which the addressee is able to construct. This approach thus focuses on the speaker's point of view, rather than the hearer's perspective.

2.6 Differences between English and German

At first glance, it seems that articles in English and German are used similarly, but there are actually many differences in their use. Rowlinson (1994: 87-90) presents a list of cases where the article is used in German but not in English and where German uses a different article. He first discusses abstract nouns (e.g. *die Eifersucht ist keine Tugend/jealousy is not virtue*[12]), genitives (e.g. *der Klang der Musik/the sound of music*[13]), months, seasons, parts of the day, meals (e.g. *der Mai ist gekommen/May is here, im Sommer nimmt man das Frühstück draußen/in summer we eat breakfast outside*[14]), and expressions of price and quantity (e.g. *sieben Mark das Kilo/seven marks a kilo*). He then adds that the definite article in German is emphatically used instead of the demonstrative *jener* (e.g. *ich möchte den Kuchen, bitte/I'd like that cake, please*), and that, with parts of the body, the definite article can be used instead of the possessive (e.g. *hebt die Hand!/put your hand up!*). Furthermore, he states that the definite article is used with some geographical names, for instance when the name of the country is feminine (e.g. *wir fahren in die Schweiz/we're travelling to Switzerland*[15]), or when the name is preceded by an adjective (e.g. *das schöne*

[12] Rowlinson (1994: 88) states that in German it is also possible to have no article, as in *das klingt wie Eifersucht/that sounds like jealousy*.
[13] Rowlinson (1994: 88) notes that spoken German might use the preposition *von*, as in *der Klang von Musik*. The article is then optional.
[14] Note that in English some of them behave differently. Compare: *It was on the radio on Sunday* and *That was the Sunday before we moved* (Biber et al. 1999: 262).
[15] Note that some geographical names in English require an article (e.g. *the USA*), others show variable article use, while others do not allow the article (e.g. *Canada, Italy, China*).

Italien/beautiful Italy[16]). Finally, in German, proper nouns can be used with a definite article, but only informally (e.g. *hast du den Günter gesehen?/have you seen Günter?*).

2.7 Variable article use in English

English articles are an area of interest for researchers, especially due to the variability in their use. But on the whole, current knowledge is largely based on standard reference grammars, and little research has been done with respect to variable article use in English. Scholars have mainly investigated this topic in the field of Corpus Linguistics and have examined it with specific lexical categories. The following provides a brief overview of the relevant studies.

Hundt (2016) uses a corpus-based and construction grammar approach to analyse diachronic change and article variation in AmE with single role predicates (i.e. *professor*, *president, governor, manager*, and *director*). Using the *Corpus of Historical American English* (COHA) as a data source and regression analysis, she finds that there is diachronic variation in article use with single role predicates in the period from 1900 to 2009. In particular, the definite article increases, while the use of bare NPs declines. In another study, Hundt (2018) uses the same methodology to investigate variable article use in present day English with a different lexical category – namely institutional nouns – comparing BrE and AmE. The data come from two different corpora: the *British National Corpus* (BNC) and the *Corpus of Contemporary American English* (COCA). While the results show an overall higher preference in AmE than BrE to use the definite article, regression analysis reveals that the head of the institutional noun is the strongest factor influencing article use. In other words, the choice of lexical item affects the use of articles more strongly than a noun's pre- and postmodification or regional variety.

Using data from the BNC, Tse (2001, 2003) investigates different grammatical factors that influence article use in multi-word organization names in British newspapers. The results of her statistical analysis (based on logistic modelling) show a clear scale of gradience between proper names, on the one hand, and common nouns, on the other. Specifically, nouns with a proper name as premodification (e.g.

[16] Note that English can vary, see for example in *das neue Deutschland/the new Germany* (Rowlinson 1994: 89).

Sheffield University) tend to omit an article, whereas nouns with a prepositional phrase as postmodification (e.g. *the Department of Trade*) are more likely to occur with an article. Her findings then confirm "'the classical' (although grossly oversimplified) assumption that common nouns require articles and proper names do not" (2003: 308). Tse (2004) makes use of the same corpus for a further analysis, which examines the presence and the omission of the definite article with personal names in present day English. In this study, she aims to give a detailed description of article use, using a grammatical perspective. The investigation includes different types of personal names: pure titles as unique references, nicknames and epithets, quasi-names given to supernatural beings, fictionalised beings, and animals treated as human beings. Tse claims that "in terms of grammatical composition and article usage, each semantic class of personal names has its own characteristics" (2004: 241). Due to the large variability through which personal names can be expressed, the usage of the definite article greatly varies.

A corpus-based approach is also used in Yoo's (2007) study, in which he explores definite article variability before *last/next time* in spoken and written AmE, investigating why, when, and how often *the* occurs. Six different corpora are used for the analysis.[17] Most strikingly, the findings reveal a stronger likelihood for *last time* to combine with *the* than *next time*. However, "the use of both *the* and Ø before *last/next time* is well-established in both spoken and written data" (Yoo 2007: 102). Additionally, in both contexts, *the* is used in nominal use and with postmodification, and Ø in adverbial use and without postmodification. More recently, in the fields of Cognitive Linguistics and Variationist Sociolinguistics, Hollmann and Siewierska (2011) use a combination of a corpus-based and Construction Grammar approach to investigate the syntactic phenomenon of the definite article reduction (DAR). In particular, they focus on the distribution of the realisations of *the* (i.e. [ðə], [ði] [t], [θ], [ʔ], Ø) in a Lancashire dialect of English. In their analysis, they take into account the phonological context, the information structure, the social value of each variable, and the token frequency. The most striking findings show that "the degree of

[17] The corpora used in Yoo's (2007) study are the following: the *Switchboard Corpus* (SWB), the *Corpus of Spoken Professional American English* (CSPAE), the *Michigan Corpus of Academic Spoken English* (MICASE), the *Santa Barbara Corpus of Spoken American English-Part 1* (UCSB), and the *UCLA Oral Corpus* (UCLA) for the spoken data, and a sub-corpus, namely the *Los Angeles Times and Washington Post* (LATWP), of a larger corpus called the *North American News Text Corpus* for the written data.

reduction (full, reduced, zero) does not appear to be phonologically conditioned" (Hollmann and Siewierska 2011: 37) and that the prepositional phrases seem to occur more often with reduced and zero articles.

Finally, some researchers (Ross 1972; Platt 1974; Huddleston and Pullum 2002; Harley 2004; Callegaro et al. 2019) have investigated article use with abbreviations. Overall, acronyms (e.g. *ACTA, EUROPOL*), being syntactically closer to proper names, tend to omit the article, while initialisms (e.g. *NLD, EU*), being syntactically closer to common nouns, are more likely to occur with an article and to be variable. In their corpus-based study, Callegaro et al. (2019) confirm the findings provided by Harley's (2004) previous study. Namely, article use of an abbreviation's definite full form is predictable based on whether the resulting abbreviation is an acronym or an initialism. More specifically, with acronyms, abbreviations tend to be bare NPs (e.g. *the North Atlantic Treaty Organisation, NATO*), while with initialisms, the abbreviations generally follow the full form (e.g. *the World Trade Organization, the WTO* and *Economic and Monetary Union, EMU*).

Regarding German, to the best of my knowledge, research has mainly examined article use and has not explored variability in detail. As described above, German articles regularly occur in front of nouns and are generally omitted in specific contexts. In other words, they are less variable (Rowlinson 1994: 90).

2.8 Article use in British English and Irish English

It is known that the use of English has steadily increased all over the world. A considerable amount of literature has been published to describe and characterize the varieties of English that have evolved throughout the centuries. Moreover, many studies have analysed the differences in article use among these varieties (see e.g. Platt et al. 1984; and Siemund 2013). As previously mentioned, the current study focuses on BrE; however, it has to be noted that the proceedings of the European Parliament, collected in *Europarl*, also include instances uttered by Irish speakers. As regards article use, IrE is markedly different from BrE (see e.g. Hickey 2007: 251; Filppula 2008: 346; Corrigan 2010: 52; and Kallen 2013: 122). More specifically, Siemund (2013: 97) attests that this variety is among those that overuse the definite

article.[18] In IrE, the definite article largely "tends to be used more than in more standard forms of English" (Hickey 2007: 251). Sand (2004: 286) explains the more widespread use of the article in IrE as a result of language contact with a language that only has the definite article, namely Irish Gaelic; therefore, "an 'overuse' of the definite article is stated and explained in terms of Irish Gaelic influence". Section 5.4.6 will investigate article use in BrE and IrE with collective nouns using data from *Europarl*.

Since this project will be based on a Construction Grammar approach, the next chapter will provide the necessary theoretical background of the Construction Grammar framework.

[18] Other varieties are Scottish, Appalachian, Newfoundland English, Singapore English, and Jamaican English.

3 Theoretical framework: the Construction Grammar approach

3.1 The Construction Grammar framework: basic concepts

The theoretical framework used in this study to analyse variable article use in English is construction-based. In the last decades, there has been an increasing interest in this new approach to grammar (see e.g. Lakoff 1987a; Langacker 1987; Sag 1997; Kay and Fillmore 1996; Ginzburg and Sag 2000; Hollmann and Siewierska 2011). The following is a brief overview of some of the basic concepts of Construction Grammar. With regard to the prospect of article use, the most important concepts addressed are: the definition of a construction and the non-compositionality aspect, the importance of semantic relations within a construction, the organization of constructions, and the various possible inheritance relations among constructions and one particular case thereof, i.e. coercion.

Similar to many other theories of grammar, the question that Construction Grammar (CxG) tries to answer is what speakers know about the language they speak. In fact, grammatical constructions "are said to reflect all the linguistic conventions that speakers of a given language know and make use of when they communicate in that language" (Fried and Östman 2004b: 23). The conceptual origins of this grammatical theory go back to Fillmore's works of the 1960s and 1980s (Fillmore 1968, 1982, 1988). Researchers agree on the fact that the perspective used by CxG strongly differs from the generative approach (e.g. Fried and Östman 2004a; Hoffmann and Trousdale 2013; Hilpert 2014). Mainstream generative theory is based on a system of rules, which separates lexicon from grammar, i.e. the lexical items (such as words or morphemes) are inserted in a finite set of grammatical principles (e.g. Cruse 2000: 238, Jackendoff 2003: 39, Taylor 2012: 19). Unlike the generative approach, Construction Grammar does not describe rules but rather constructions, in which "[l]exicon and grammar are not distinct components, but form a continuum of constructions" (Langacker 2005: 102). In the constructional view, the basic assumption is that "[t]he totality of our knowledge of language is captured by a network of constructions: a 'construct-i-con'" (Goldberg 2003: 219). An early definition used to describe a construction comes from Goldberg (1995: 4) and reads as follows:

> C is a CONSTRUCTION iff$_{def}$ C is a form-meaning pair <F_i, S_i> such that some aspect of F_i or some aspect of S_i is not strictly predictable from C's component parts or from other previously established constructions.

Contrary to the compositionality conception of generative grammar (i.e. lexical items are put together by grammatical rules), Construction Grammar suggests non-compositionality. The idea that semantics has to be compositional dates back to Frege (1967) and was further developed by Montague (1974). The Principle of Compositionality assumes that "the meanings of individual words can be used to build up the meanings of larger units: the meaning of the whole is determined by the meaning of its parts […]" (Crystal 2008: 96). Constructionists reject this principle because compositionality cannot then explain idiomatic expressions (e.g. *kick the bucket*, *break a leg*, *break the ice, speak of the devil*, or *hang in there*), whose meanings are not inferred from the meanings of their parts. Therefore, an advantage of Construction Grammar is that the Principle of Compositionality is not needed because the constructional meaning is given by the construction itself and not exclusively by the individual lexical items.[19]

There are, however, many expressions, such as *I love you* or *I don't know*, that are semantically and structurally clear and whose meaning is predictable and compositionally derived (Hilpert 2014: 13). According to the early definition by Goldberg (1995: 4), these expressions might not then be seen as constructions. Therefore, a revision of the initial construction definition was needed. In the more recent accounts of the constructional framework, the reason why these are considered constructions all the same is the fact that they are frequently used by speakers. As Hilpert states (2014: 13), "[s]ome expressions may superficially look like constructs, but through repeated use, they have become the default option for a specific communicative situation." In other words, speakers are constantly able to understand novel constructions and are also capable of learning expressions as forms, based on their high usage frequency. In order to include this aspect in the definition of a construction, Goldberg (2006: 5) suggested the following updated version:

[19] Note that, for instance, in *Cognitive Grammar* Langacker (2005: 140) is "against the traditional assumption of full compositionality for sentential semantics", but also criticizes the non-compositionality approach of Construction Grammar as well, arguing that it is not strictly necessary as compositional units can equally be "psychologically entrenched and conventional in the speech community."

> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency.

In the previous description, these forms would have been simply analysed as constructs (e.g. *I love you* would have been seen as a basic example of the transitive construction). On the contrary, the newer definition is extended and considers also those expressions that occur frequently enough to be stored in the construct-i-con as forms. What remains between the first definition and the revised one is the notion that constructions are form and meaning pairings. Croft (2001: 18), in his Radical Construction Grammar account, proposes the following symbolic structure of a construction, illustrated in Figure 3.1.

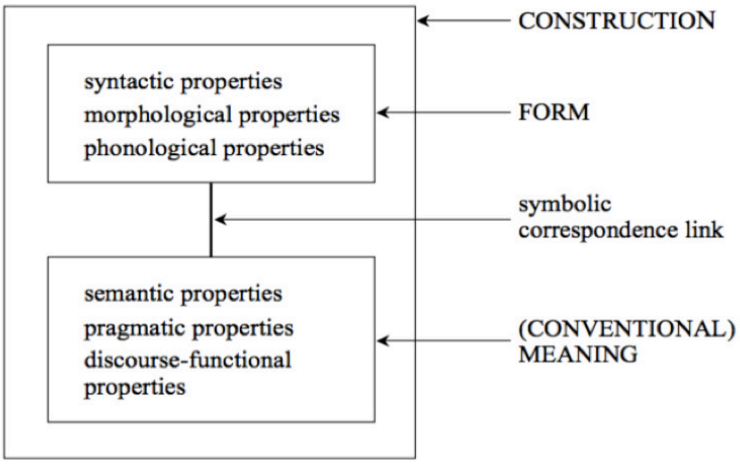

Figure 3.1: Radical Construction Grammar representation of a construction (Croft 2001: 18).

The outer box represents a construction and includes two smaller boxes, which refer to form and (conventional) meaning. The former includes its syntactic, morphological and phonological properties, while the latter involves semantic, pragmatic and discourse-functional properties. According to Croft (2001), form and meaning are connected via a symbolic link, which is internal to the construction.

The extended diagram of the internal structure of a construction is given in Figure 3.2 below. One can note that the syntactic structure contains formal elements, but there are no syntactic relations that connect these elements of the construction. By contrast, the semantic structure is even more complex because it "consists of both the

components of the semantic structure and the semantic relations that hold between the components of the semantic structure" (Croft 2005: 12). Thus, to be complete, a construction's representation must include the relations between the syntactic structure's elements with the semantic structure's components, i.e. symbolic relations. If the correspondences between the elements and the components were not available, a hearer would not be able to understand the meaning of an utterance based on its form. Therefore, according to Croft (2005), syntactic relations (i.e. the relations between the elements of the syntactic structure) are not essential for communication. To understand a speaker's utterances, a hearer only needs to identify three factors, namely the form of the construction, its meaning, and the correspondences between the syntactic elements of the construction and the components of its semantic structure. In other words, syntactic relations between elements do not exist.

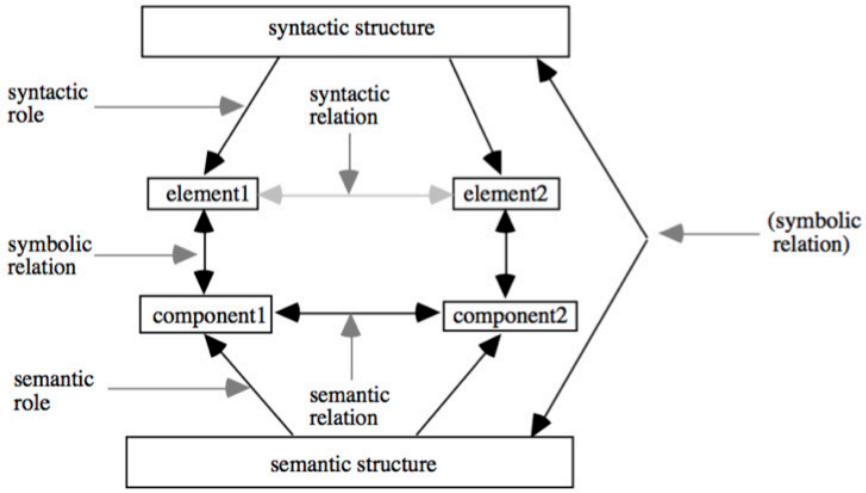

Figure 3.2: Expanded diagram of a construction's internal structure (Croft 2005: 12).

Of great importance are the semantic relations discussed by Croft (2001, 2005), as they will be the focus on which the constructions presented in section 6.5 are based. The constructional representations in relation to article use will not take into account the syntactic relations between the elements; they will rather approach the entire analysis from what semantic properties are present in the elements within an NP and what semantic properties they transfer onto the whole construction.

A further point on which Radical Construction Grammar is based is the non-reductionism aspect. Crucially, in reductionist theories, complex phenomena are

described in terms of their primitive constructs (i.e. core constituents), which in turn cannot be classified into smaller components. By contrast, in Radical Construction Grammar, primitive constructs are the constructions themselves, which are already complex structures. In Croft's (2001, 2005) view, constructions then exist at every level of language (i.e. from the phoneme up to the sentence level via morpheme, word, phrase, and clause). Thus, grammatical constructions "are symbolic signs and represent the basic building blocks of linguistic analysis" (Fried and Östman 2004b: 18). Words, for instance, by being symbolic form and meaning pairings, are also constructions, as are affixes, idioms, syntactic patterns, and their interactions. In other words, all constructions are "associated with more or less detailed information about [their] phonological, morphological, syntactic, semantic, pragmatic, discourse, and prosodic characteristics" (Fried and Östman 2004b: 13). A construction is, then, "an *abstract*, representational identity, a conventional pattern of linguistic structure", whereas "the actually occurring linguistic expressions, such as sentences and phrases" are defined as its constructs (Fried and Östman 2004b: 18, emphasis original).

With respect to the organization of constructions, other scholars have formulated the theory that constructions can further be analysed in terms of different levels of abstraction. Traugott (2008: 236) and Trousdale (2008: 169) have suggested the following hierarchical ordering:

i. *macro-constructions*: pairings of form and meaning defined by structure and function (e.g. Partitive Construction, Degree Modifier Constructions);
ii. *meso-constructions*: a series of specific constructions that behave similarly;
iii. *micro-constructions*: individual types of constructions;
iv. *constructs*: the actual tokens used in a construction that can be replaced by other constructs.

As pointed out by Traugott (2008: 236), the constructional hierarchy is not limited to these levels. Indeed, some phenomena might need more complex and elaborated levels of generalizations, while others might require fewer levels.[20] Therefore, every

[20] The organization of constructions along different levels of abstraction recalls the concept of schematicity discussed in Cognitive Grammar; the nodes of the constructional network in fact "include complex expressions as well as constructional schemas, characterized at various levels of schematicity and linked by categorizing relationships of elaboration and extension" (Langacker 1999: 20).

grammatical construction is developed on a vertical network (i.e. from a more abstract level on the top to a more concrete level at the bottom) and can be analysed on various levels of abstraction.

This constructional hierarchal system is linked to another important property of Construction Grammar, namely the concept of inheritance relations, which does not focus on one construction but rather relates to the nodes existing between two or more constructions. Fried and Östman (2004b: 23), for instance, claim that inheritance "provides a coherent way of capturing which properties individual constructions have in common and what sets them apart as related, but distinct to grammar patterns." Put differently, the general idea is that "a construct can be an instance of multiple types at once" (Michaelis 2013: 144). According to Hilpert (2014: 57-58), constructions of different types share information and are situated on a vertical continuum. Similar to the constructional organization suggested by Traugott (2008) and Trousdale (2008), more abstract constructions are located on top of the constructional network, while more specific and concrete constructions are found towards the bottom. In other words, the characteristics of the constructions (i.e. the characteristics of form and meaning) are "inherited in a downwards directions, from higher, more schematic levels towards lower, more concrete levels" (Hilpert 2014: 58). Goldberg (2013: 21-22) explains what she calls a default inheritance network (i.e. inheritance relations) by focussing on the following examples: *in prison*, *from school*, *for work*, *on vacation*, *to bed* and *in hospital*.[21] What these phrases have in common is the fact that they combine a preposition with a bare count noun, i.e. the P N construction.[22] Goldberg (2013) states that this construction is a basic construction which derives from a more abstract and general construction, namely the PP construction (i.e. the prepositional phrase construction). More specifically, the P N construction inherits a specific feature from the PP construction, that is, its word order (i.e. the preposition is always located in front of the noun). However, as pointed out by Goldberg (2013: 18), the P N construction is considered unusual because prepositional phrases with common count nouns usually require a determiner and allow for premodification (*She went to the big bed* vs. **She went to big bed*).

[21] Note that *in hospital* is normally used in BrE.

[22] Note that, in the P N construction, Goldberg (2013: 13) does not see a bare NP as a noun phrase but as a noun, i.e. the underlying slot changes from NP to N along the constructional hierarchy. In the current study, contrary to Goldberg (2013), a bare NP will be considered an NP, i.e. a more detailed definition of the NP slot.

Furthermore, contrary to the prepositional phrase construction, the P N construction cannot be formed with any type of nouns (*go to bed* vs. **go to couch*). Thus, the P N construction is not productive but rather limited. Hence, inheritance refers to "a 'downwards' relation; more specific constructional characteristics are projected 'upwards'" (Hilpert 2014: 59).

Based on this argumentation, it could be concluded that the exchange of properties between constructions merely happens along a hierarchal scheme (i.e. a vertical structure). There are, however, various types of inheritance networks. The one discussed above is the most basic one and has been termed *instance link* by Goldberg (1995: 79); in this type of network, one construction is considered as a special case of another construction and is therefore a more specific version, i.e. inheritance happens from one level of abstraction to another. A second type concerns *polysemy links* (Goldberg 1995: 75), which "capture the nature of the semantic relations between a particular sense of a construction and any extensions from this sense." A good example is the English S-Genitive construction. The central meaning of this construction is *possession*, which "is related to extended senses of the construction via polysemy links" (Hilpert 2014: 61). For instance, the expression *John's book* denotes a type of possession that differs from the one expressed in *John's train* or *the country's president*. In the first case, the book is owned by John, while in the second case, John is not the owner of the train and the president does not own the country, the emphases are put on the train John is travelling on and the country that the president is governing. Therefore, the meaning expressed by *John's train* and *the country's president* are in turn to be considered as an extended sense of the primary meaning. The extended meanings are therefore connected to the main construction's meaning via polysemy links. This type of inheritance is similar to a third kind that Goldberg (1995: 81) identifies as *metaphorical links*. The difference between this type and the previous one is that the senses here are connected to each other via a conceptual metaphor. To illustrate, the meaning of the caused motion construction (e.g. *Pat threw the metal off the table*) is linked with the meaning of the resultative construction (e.g. *Pat hammered the metal flat*) via the *change is motion metaphor* (Goldberg 1995: 81). In the former construction, the metal moves physically in the space (i.e. from point A to point B), while in the latter, the movement refers to a change of shape (i.e. from non-flat to flat). The two constructions are therefore conceptually interconnected through a metaphorical link. A fourth type of inheritance is defined as *subpart link*

and happens when "one construction is a *proper subpart* of another construction and exists independently" (Goldberg 1995: 78, emphasis original). In other words, two constructions are related to each other because of a formal or semantic overlap; however, this does not mean that the second construction is the first construction's daughter; on the contrary, they are two autonomous constructions. To cite an instance, the transitive construction (e.g. *John wrote a letter*) and the ditransitive construction (e.g. *John wrote Mary a letter*) are independent constructions but share some features, namely the subject with an agent role and the direct object with a patient or theme role (Hilpert 2014: 62). Therefore, contrary to the previous kinds of inheritance that link different levels of abstraction in the constructional network, subpart links connect constructions that are situated on the same level of abstraction. As Hilpert (2014: 63) claims, "[r]ather than one construction linking to just one other construction, the construct-i-con is thus a network with many-to-many links." It is for this reason that subpart links are closely related to the concept of *multiple inheritance*, in which an instance can instantiate two (or more) different abstract constructions at the same time (Goldberg 1995: 97, Michaelis and Lambrecht 1996: 237). An example is the following sentence, which is discussed by Hilpert (2014: 63-64).

(1) The Smiths felt it was an important enough song to put on their last single.

The instance contains two constructions that are linked and combined together. The noun phrase *an important song* belongs, for instance, to the attributive adjective construction. This is blended into a second construction, i.e. the enough to-infinitive construction, which usually consists of a phrase whose phrasal head is modified by *enough* and followed by the preposition *to* and an infinitive clause (e.g. *You're old enough to know better*). In (1), the noun phrase is followed by a *to*-infinite clause, but *enough* modifies the adjective *important* and not the phrasal head. These two constructions are thus intermixed and connected via multiple inheritance relations.

At this point, it is worth highlighting a further central concept of Construction Grammar related to multiple inheritance, i.e. the principle of coercion. This phenomenon describes cases where a word can change its meaning based on the construction it is in. In other words, "the meaning of a lexical item may vary systematically with the constructional contexts in which it is found" (Hilpert 2014: 17). Coercion, therefore, allows phrases to have "conventionalized alternative

interpretations" (Jackendoff 2013: 82). Examples of coercion effects are given in (2) and (3).[23]

(2) Three beers please!
(3) Give me a butter.

In their primary sense, *beer* and *butter* are mass nouns and, therefore, could not be used in the plural, nor with the indefinite article. In the examples above, however, *beer* follows a numeral and is inflected in the plural form, while *butter* is combined with the indefinite article. Hence, these lexical items change reference and pass from mass nouns to count nouns. This particular nominal coercion phenomenon is defined as a mass – count coercion.[24] Put differently, *beer* and *butter* have been coerced by the new construction and, in turn, change the meaning and have a related interpretation. However, it has to be noted that lexemes cannot be coerced into receiving any possible interpretation. As Goldberg (1995: 159) affirms, "[i]n order for coercion to be possible, there needs to be a relationship between the inherent meaning of the lexical items and the coerced interpretation." Coercion effects are then the results of successful inheritance relations between constructions.

To conclude this section, constructions should not be seen as a basic constructional hierarchy but rather as a complex network of constructions, connected to each other via various links at the same time. In recent years, there has been an increased interest in the constructionist approach. This has led to a proliferation of studies in several research areas (e.g. corpus linguistics, language variation and change, typology, language acquisition, language processing, and computational applications). However, further work is required in linguistic research and between linguistics and other disciplines (Goldberg 2013: 31). The following pages will present what construction grammarians have discussed with regard to English articles and will give an overview on how Construction Grammar has been applied in language variation.

[23] Example (2) is discussed by Hilpert (2014: 17), whereas example (3) is discussed by Michaelis (2013: 148).
[24] In Construction Grammar literature, the terminology referring to this type of construction may vary. Hilpert (2014: 17), for instance, calls them individuation constructions, while Michaelis (2013: 148) talks about derivational constructions.

3.2 Construction Grammar and article use

In "The Mechanisms of 'Construction Grammar'", Fillmore (1988) adopts a boxes-within-boxes notation to describe constructions' external and internal syntax (i.e. the features of a construction as a whole and the constituents it consists of). The external structure is represented by a large box, which includes smaller boxes containing the internal specification of the construction. The structure as a whole marks the dominance relationships (i.e. the unification relations) between the elements of a construction. In addition, sets of bracketed attribute-value pairs are used to represent the grammatical information of a construction. An attribute denotes a particular property (i.e. syntactic, semantic, or pragmatic category), while the value gives more specific details of that property. A value can be either binary (e.g. definiteness: +/–, maximality: +/–) or non-binary (e.g. lexical category: N, Adj., V, etc.). Both binary and non-binary values can also be 'unspecified'; in this case, the brackets are left empty. Fillmore (1988: 39-41) offers a first attempt at defining the determination construction, which is used in his work to explain how the constituents of a construction are unified, satisfying the structural positions' requirements. Figure 3.3 shows the representation of Fillmore's determination construction.

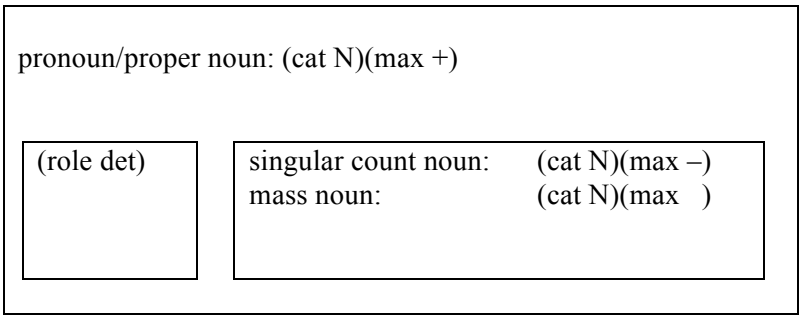

Figure 3.3: Structure of the determination construction (adapted from Fillmore, 1988: 40).

When analysing headed constructions, Fillmore (1988: 38) subdivides them into different levels and distinguishes between *maximal* and *minimal categories*; the former "fill major structural positions in constructions", whereas the latter "are the stored or derived units of the lexicon." Major category units are pairs of categories' features and level types. For instance, a maximal noun phrase is expressed as (cat N)(max +), while a lexical adjective as (cat A)(min +). According to Fillmore's (1988: 39) distinction, the determination construction "consists of a maximal noun

phrase containing a determiner and a non-maximal nominal head. As shown in Figure 3.3, the determination construction is composed of two slots. More specifically, the box on the left contains the determiner, while the box on the right can be filled by either a non-maximal nominal (i.e. singular count noun) or a nominal whose maximality value is not specified (i.e. a mass noun). It is important to highlight the fact that the determiner role is not filled exclusively by articles; possessives and demonstratives can also appear in the determiner slot. The combination of these two slots represents a maximal noun phrase. Since the maximality value of pronouns and proper nouns is marked positively, they do not require a determiner. Put differently, in Fillmore's (1988) structure, the outer box can be seen as a container which needs to have a fixed maximality: either this is filled by a pronoun or a proper noun that are themselves maximal or the lacking maximality of other nouns needs to be complemented by a determiner completing the maximality requirement of the box.[25]

The structure of the determination construction is considerably basic, but noun phrases are often more complex than this representation suggests. As Fillmore (1988: 40) points out, the diagram should include "the morphology for de-marking count nouns when they are plural" and reflect that "both mass nouns and proper nouns have special uses in which they exhibit the syntax of count nouns." Furthermore, he (1988: 40-41) claims that there are other constructions in English in which a non-maximal nominal does not require the 'obligatory' determiner, such as the unique-role nominal predicate construction, as in (4), and the fronting-to-*that* construction, as in (5).

(4) She is chief surgeon to the royal family.
(5) Foolish child that I was.

Every position within a construction must have specific features in order to make the whole construction work successfully. Thus, in a determination construction, number and definiteness are properties that naturally become an essential part of the construction itself. For instance, the expression **these butter* is not possible because the plural demonstrative cannot be combined with a mass noun. Therefore, a maximal

[25] This view seems to tie in with the idea – strongly supported in Generative Grammar – of having the head of a noun phrase being a determiner, making it a determiner phrase (i.e. *DP*). As pointed out by Croft (2001: 257-258), generative grammarians claim that the noun phrase is not full (i.e. it cannot fulfil its referential function) without the presence of the determiner. This is similar to Fillmore's (1988) structure, which also implies that an NP is not 'complete' without a determiner.

nominal needs "to be recognized as singular or plural, […] and as definite and indefinite, establishing its qualification for inclusion in certain of the existential sentence constructions" (Fillmore 1988: 41). Thus, unification processes (i.e. the combination of categories with matching values) are fundamental for grammatical agreement, which is essential for successfully completing the NP. Therefore, in order to satisfy the needs of the NP, the maximality value has to be respected, as well as the connecting features between the noun and the determiner.

The latter point is amply discussed by Fried and Östman (2004b), who analyse in more detail the specific properties that should be combined in a determination construction.[26] Figure 3.4 presents the lexical construction of the definite article *the*. In the box, the syntactic attributes (*syn*) include the lexical category (*cat*), i.e. article (*art*), and the phrasal/non-phrasal attributes (*max* and *lex*). These must be combined with the corresponding semantic properties (*sem*), whose attributes, namely *frame*,[27] configuration (*cnfg*), number (*num*), and boundness (*bounded*), are, in this case, unspecified.

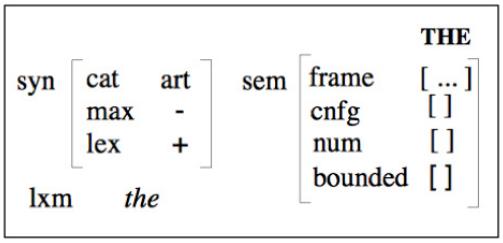

Figure 3.4: Representation of the determiner *the* (Fried and Östman 2004b: 33).

When the definite article is combined with a noun, the construction changes and the semantic properties of the determiner are not unspecified anymore but match the features of the noun it is related to. Practical examples are provided in Figure 3.5, which shows the representations of the constructs *the snow* and *the book*. In the first case, the lexical item *snow* is a mass, singular and unbound noun; therefore, the semantic properties of the definite article inherit the same characteristics. On the other

[26] In their analysis, Fried and Östman (2004: 33-36) also include the quantifier *much* and compare the representations of the constructions *the snow/the book* vs. *much snow/*much book*.

[27] The notation […] means that the value needs to be specified or is not explained for space reasons or because it is not the focal point of the construction.

hand, *book* is a countable, singular and bound noun and the definite article takes the same semantic properties. In other words, similar to Fillmore's (1988) argumentation, we see that a noun phrase has at least two constituents, i.e. a determiner and a noun, whose syntactic properties are specified in the corresponding boxes (see for instance Figure 3.4), while the semantic features are "a union of the two frames associated with the two constituents" (Fried and Östman 2004b: 36). For both Fillmore (1988) and Fried and Östman (2004b), within an NP, the noun dominates the construction, i.e. the noun determines whether there is the need for a determiner to complete the maximality requirement of a noun phrase, and the noun gives the features to the article.

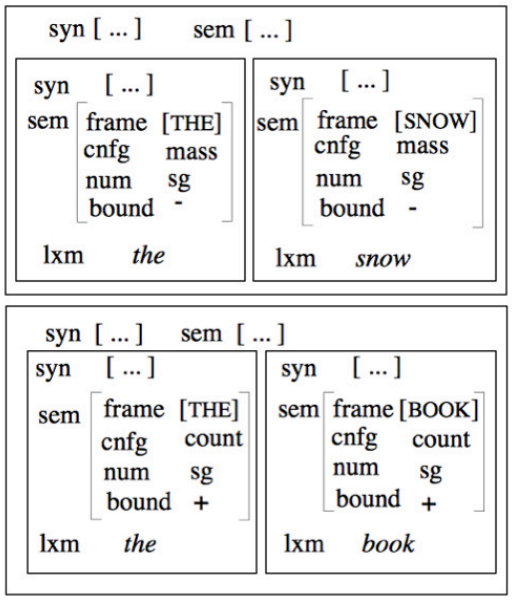

Figure 3.5: Representation of the constructions *the snow* and *the book* (Fried and Östman 2004b: 35).

It is worth noting that constructions have external properties that need to be considered as well and that a phrase is a complex net of relationships between the constituents and their attributes. One of the advantages of CxG is that it does not only show which constructs can instantiate a construction and how they are combined, but it also allows us to create generalizations on a more abstract level. Fried and Östman (2004b: 36-37) therefore step back from the specifications of the construction's lexical items and suggest their abstract generalization of the English Determination construction; its representation is shown in Figure 3.6.

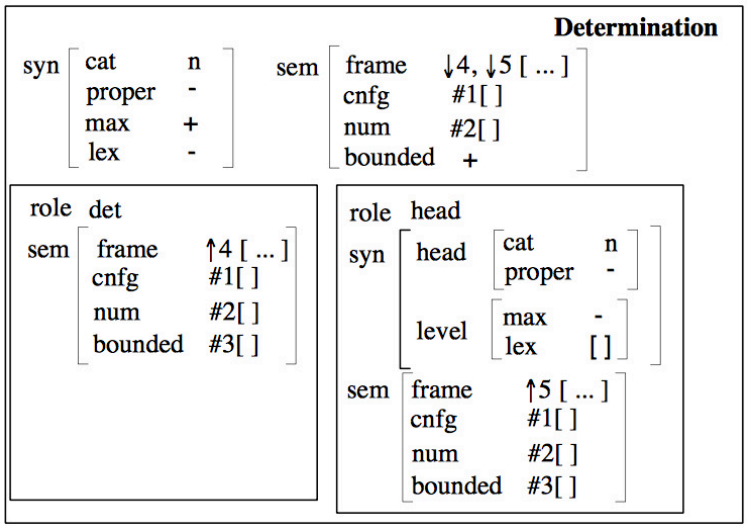

Figure 3.6: Representation of the English Determination construction (Fried and Östman 2004b: 37).

The following are some significant points that need to be considered when interpreting the representation of the construction. First, the general representation of the determination construction in English does not only focus on the definite article but also takes other determiners into account (e.g. *a(n)* and *much*). Second, this construction accurately allows for expressions such as *the book*, *a book*, *the books*, *much snow* and excludes others such as **a snow*, **the snows*, **a books*, **much book(s)*, and **the Prague*. Third, since the lexical value of the head is not specified (*lex [ ]*), a noun can be preceded by a modifier (e.g. *the beautiful book*). Finally, the configuration and number values of the external semantic properties are unspecified because they depend on the corresponding values of the constituents. Generally, in order to have a successful construction, the "unification can take place only on condition that the relevant pieces of information do not conflict"; in other words, "two values either have to match exactly or at least one must be unspecified" (Fried and Östman 2004b: 38). For example, in the above expression **a books* the unification is not possible, as the number value of the noun *book* is plural and that of the indefinite article is singular, whereas *the books* would be possible because the number value of the definite article is unspecified and can take on the value of the noun, i.e. they unify.

Fillmore (1988) and Fried and Östman (2004b) provide a good but too broad analysis of the English determination construction. Besides the previously mentioned

corpus studies conducted by Hundt (2016, 2018, see section 2.7), to the best of my knowledge, as of yet, no other detailed corpus-based investigation on article use has been done within the construction grammar framework, nor has its variability been taken into account. Both Fillmore (1988) and Fried and Östman (2004b) acknowledge that their representations might have some limitations. These representations, in fact, suffer from some serious drawbacks.

Firstly, Fillmore's (1988) diagram includes articles, demonstratives, and possessives, while Fried and Östman's (2004b) representation combines articles with quantifiers. With respect to article use in CxG, a very important question that might be asked is whether including all determiners within the same construction (i.e. the determination construction) is the best approach. Fillmore (1988) and Fried and Östman (2004b) do not clearly distinguish between the determiners and their equivalent meanings (e.g. *this apple* differs from *the apple*, which, in turn, has a different meaning from *an apple*). A much more systematic approach would examine the determination construction by separating the different types of determiners. Fillmore (1988) and Fried and Östman (2004b) do not ascertain that their representational view may be the highest level of abstraction of the determination construction and that there might be more specific constructions at lower abstraction levels (i.e. along the constructional network). Grammatical constructions such as the article constructions, the demonstrative constructions, the possessive constructions, and the quantifier constructions might therefore be situated on the same level of abstraction and share some characteristics between each other and with the determination construction, too.[28]

A second limitation refers to Fillmore's (1988) diagram, which, after more recent additions by other scholars, can be considered too simplistic, as it does not include the inheritance relations between the noun phrase's constituents. Contrary to Fillmore (1988), Fried and Östman (2004b) pay particular attention to the concept of inheritance. However, since the noun determines the whole construction, they state that only the determiner can inherit the characteristics of the noun and do not address the question whether the noun might inherit some features from the determiner, too.

[28] This view is shared by Trousdale (2008: 169-170), who points out that the possessive and the demonstrative constructions are meso-level constructions and are instances of the determiner construction, which represents a higher level of abstraction and is located at the macro-level along the constructional hierarchy.

As seen previously, the definite article, for instance, conveys a different meaning from the indefinite article: *the book* differs from *a book* even though the noun does not change the category; in both cases, *book* is a concrete count noun. Likewise, as discussed in Chapter 2, the way articles are used determines whether an NP has a specific or a generic reference. Therefore, what Fried and Östman (2004b) suggest is that inheritance relations only take place in the direction from the noun to the article. However, this *unilateral* inheritance relation[29] proposed by Fried and Östman is not sufficient in accounting for differences in NPs that contain different types of determiners.

Furthermore, another weakness regards the types of nouns both Fillmore (1988) and Fried and Östman (2004) consider in their representations. With respect to the nouns occurring with an article, they only focus on singular count nouns (e.g. *lawyer, apple, building*) and mass nouns (e.g. *snow, water*). However, in the case of the nouns that do not require an article, they exclusively mention proper nouns (e.g. *Prague, Italy*). They thus fail to acknowledge that variability involves more than these noun types. Abstract nouns, for instance, are not proper nouns and do not normally require an article. They do not give consideration to all those cases in which article omission is involved, nor to the possibility for articles to be variable. Hence, there is a need to develop more narrow and specific representations that only focus on English articles and allow for contexts in which articles do not occur and/or are variable.

As previously indicated, since constructions are linked to each other through a complex constructional network, it is important to note that a construction can have external properties interplaying and influencing the NP construction itself. As far as article use is concerned, one aspect that needs to be remembered is its anaphoric reference and the resulting distinction between new information (usually expressed with the use of the indefinite article) and given information (usually expressed with the use of the definite article). In other words, noun phrases can have a wide range of meanings that can change based on the discourse meaning, e.g. whether an NP was already mentioned before or not. However, this aspect is not taken into consideration

[29] Understanding that inheritance is always one-directional by nature, the expression *unilateral* is used for the sake of simplicity to refer to the idea that inheritance within the whole construction exclusively happens from one element to the other, i.e. in this case from the noun to the article, whereas *bilateral* inheritance relation will be suggested to refer to the possibility for inheritance also to take place in the other direction (although technically being two unilateral inheritance relations in opposite directions).

but is definitely a relevant issue for future research. The main focus of the current analysis is on English article use within an NP, and one of the main aims is the investigation of the interrelations of elements within the construction and the influence of different noun types.

3.3 Construction Grammar and language variation

Variation is a linguistic process that naturally happens in any living language. Linguistic variation has received significant attention by scholars, and a considerable amount of literature has been published in this field. In particular, this research area finds its roots in the long tradition of quantitative sociolinguistics (see e.g. Labov 1969, 1994, 2001; Tagliamonte 2006; Trousdale 2010). However, as Hilpert (2014: 185) points out, "[t]he analysis of linguistic variation has only recently been put on the research agenda of Construction Grammarians, who are thus relative late-comers" to this topic. As mentioned before, the main goal of Construction Grammar is to understand what speakers know about the language they speak. Therefore, one of the most important aspects in language variation in relation to CxG is to find out "how speakers choose between alternative constructions" (Hilpert 2014: 191).

Variation in constructions regards the possibility for a speaker to prefer a construction to another. Hoffmann and Trousdale (2011: 9), for instance, claim that "[f]rom a usage-based construction grammar perspective it is their mental construction network that allows speakers to make these choices". More specifically, in synchronic variation, they (Hoffmann and Trousdale 2011: 7) state that

> whenever speakers can choose between alternative structures there are linguistic as well as social factors that systematically affect the choice of a particular variant. In quantitative language variation parlance, the choice of a particular variant of a dependent variable is influenced by independent factors such as its linguistic context, the stylistic level of the discourse and social characteristics of the speaker.

Therefore, the generalisations found in constructional variation "are not quite as simplistic as a one-to-one mapping of a single, invariant form to a single, invariant meaning"; on the contrary, "both the formal pole and the meaning pole of a construction should be seen as containing information on several variants – formal variants of the construction as well as meaning variants" (Hilpert 2014: 181). Speakers thus know that, for instance, the indefinite article in the same construction

can be realised in two variants – i.e. *a* or *an* – without change in meaning. Variability in CxG then tells us that linguistic generalisations (i.e. constructions) "are not fixed schematic templates, like assembly instructions that allow only a single correct way of constructing a complex whole" (Hilpert 2014: 185). In particular, when analysing variation in constructions, Hoffmann and Trousdale (2011: 3) note that there are three "properties associated with variation: patterns of structural variation; the context in which the variation occurs; and the statistical correlates of frequency of use." They highlight the factors that influence one variant or construction to another, namely frequency, processing, preemption, and motivation. Speakers generally tend to use the construction with a "high type frequency, that is, those that have been encountered with many different lexicalizations […], all of which share a common meaning" (Hoffmann and Trousdale 2011: 5). Moreover, when two constructions compete against each other, speakers have the tendency to favour the simpler one, which will in turn increase the frequency with a bigger range of lexicalizations (i.e. processing effect). Preemption also plays an important role: when a speaker prefers a construction to another, the hearer automatically assumes that this has a different function from the alternative(s). This differentiation will then bring the hearer to associate every variant with specific linguistic and social contexts. In other words, "preemption encourages originally synonymous constructions to be interpreted as contextually-determined variants" (Hoffmann and Trousdale 2011: 6). In the long run, therefore, the consequence of this process might be diachronic change. Finally, motivation is a significant factor: "the more closely two constructions are related semantically, the more related they will be formally" (Hoffmann and Trousdale 2011: 6-7).

Overall, these observations show that speakers have knowledge of variation because they constantly have to make linguistic choices which depend on both linguistic and external factors; they know when to use which variant and in what context. The linguistic knowledge of the speakers thus includes the capability to distinguish between various constructions. When discussing variation in Construction Grammar, we therefore refer to a specific level of abstraction where constructions can be used interchangeably; hence, that level of abstraction is subject to this kind of variability. Based on the hierarchical system suggested by Traugott (2008: 236), (variable) article use in English will be regarded as variation on both the micro-level

and meso-level, due to the intention to move from construct level corpus findings to more abstraction. This aspect will be discussed in more detail in Chapter 6.

As mentioned in the previous section, on the question of variable article use in English, the aim of this thesis is to evaluate variability with a particular focus on bare NPs and to find novel cases in which articles might be used variably. From this, the following research questions can be formulated:

1. Are there cases of article variability at NP level based on corpus evidence?
2. How can we represent a revised model of individual NP constructions?
3. What does an evidence-based CxG construct-i-con look like that takes into account article variability?
4. Is there evidence of influence on article variability beyond the NP to guide future research?

In the chapter that follows, the parallel-corpus approach taken to investigate variable article use will be presented and the corpus this project is based on will be introduced.

4 Data

4.1 (Parallel) Corpora and the used corpus

The present study is a corpus-based project and the evidence used for the analysis thus "derive[s] directly from text" (Kennedy 1998: 7). Thanks to computers, data storage, and technological innovations, the past 50 years have seen increasingly rapid advances in the field of Corpus Linguistics. According to Church and Mercer (1993: 1), the great quantity of available data might be one of the "most immediate reason[s] for this empirical renaissance." Corpus linguistics examines "the behavioural manifestation of language, in the form of naturally-occurring spoken and written discourse" and focuses on "linguistic performance rather than linguistic competence" (Leech 1992: 107). A corpus-based approach has made it possible to analyse and describe language use from a new perspective (Biber, Conrad and Reppen 1998: 233; Kennedy 1998: 204; Partington 1998: 1; Tognini-Bonelli 2001: 2). More recently, further developments in the field have also led to a new research interest, namely parallel corpus linguistics (Borin 2002: 1). There, the focus is not on a collection of texts in the same language, but rather on "texts in one language, or language variety, together with corresponding texts in another language" (Borin 2002: 4), i.e. *parallel corpora*. Such corpora are a remarkable tool for language analysis and contribute to the advancement of many linguistic areas, such as contrastive, comparative, typological, grammatical, and lexicographical studies, language teaching, and machine translation research (Greenbaum 1996: 10; Borin 2002: 14). In particular, the translation relation existing between the parallel texts has attracted a great deal of interest in translation studies. For instance, Sinclair (1992: 395) claims that:

> The new corpus resources are expected to have a profound effect on the translations of the future. Attempts at machine translation have consistently demonstrated to linguists that they do not know enough about the languages concerned to effect an acceptable translation. In principle, the corpora can provide the information.

Parallel corpora offer an important contribution to linguistic investigation because they can compare original texts and translations across languages, original texts across languages, translations across languages, and original texts and translations within the same language (Johansson 2002: 48). More specifically, Zanettin (1998:

616-617) classifies three different types of corpora that have been used in translation studies: a) the *multi-source-language monolingual comparable corpus*, which consists of an original text and similar texts translated into the language of the original text; b) the *bilingual (or multilingual) corpus* (a parallel corpus in the narrow sense), in which language pairs are put in contrast. This type of corpus usually consists of an original text and its equivalent translation, and the relationship between the two texts is one-directional, i.e. from the source language to the target language; and c) the *bilingual comparable corpus*, in which various texts of different languages are collected and put together in respect to the content, type, and function. Zanettin (1998: 627) indicates that a "contrastive analysis of comparable corpora can reveal how similar ideas and concepts are expressed in similar texts in different languages". In other words, parallel texts constitute a valuable tool because they allow for the detection of similarities and/or differences between two or more languages. In general, there is an increasing awareness that corpora contribute to the improvement and progress of translation studies, and, as Øverås (1998: 558) states, "corpus projects of various kinds are encouraged because they facilitate comparison of series of texts or translation problems." On the whole, parallel corpora are therefore useful "to learn more about language in general" (Johansson 2007: 316).

4.1.1 Parallel corpora for the investigation of (variable) article use

In the present project, the parliamentary transcripts collected in the *Europarl Corpus* (discussed in more detail in section 4.1.2) form the basis of analysis. For the investigation of article variation, in particular, the parallel nature of the parliamentary debates' transcriptions are very useful because they permit us to focus on the contexts in which noun phrases appear without an article. More specifically, it is possible to take a language that is known to use articles frequently as a starting point and to then retrieve aligned parallel instances of a second language, which is likely to show article variability. Therefore, parallel texts provide us with the possibility to obtain instances, in which one language – i.e. German in this case – uses an article, while in the other language – i.e. English – an article is not required or is variable.

There are several reasons why German was chosen as the *mirror* language. Firstly, as already discussed in Chapter 2, English and German share a similar article categorization; that is, the differentiation between definites (i.e. *the* in English and

der, *die*, and *das* in German) and indefinites (i.e. *a*/*an* in English and *ein* and *eine* in German). Secondly, German articles frequently occur in front of nouns (*Duden* Online 2017), and a difference in frequency between German and English is expected to be found. The experimental methodology adopted in the current study, as well as the great advantages provided by parallel corpora, enable us to examine bare contexts, which are known to be very difficult to access in monolingual corpora.

4.1.2 The Europarl Corpus and CoStEP

Europarl is a parallel, sentence-aligned, and part-of-speech tagged corpus. It includes parallel texts in the following official languages of the European Union: Romance languages (i.e. French, Italian, Spanish, Portuguese, and Romanian), Germanic languages (i.e. English, Dutch, German, Danish, and Swedish), Slavic languages (i.e. Bulgarian, Czech, Polish, Slovak, and Slovene), Finno-Ugric languages (i.e. Finnish, Hungarian, and Estonian), Baltic languages (i.e. Latvian and Lithuanian), and Greek. The transcriptions of the parliamentary debates as well as their translations are freely available on the European Parliament website.[30] The earliest texts date back to 1996 and cover a time span of 15 years. The corpus "comprises of [sic!] about 30 million words for each of the 11 official languages of the European Union" (Koehn 2005: 79). Since more countries have joined the European Union, more languages have gradually been added and documents have in turn been translated into more languages.

Due to several errors found in the corpus (e.g. coding, orthography, missing data, and processing), a cleaned and improved version was developed in Zürich, namely *CoStEP – Corrected & Structured Europarl Corpus* (Graën, Batinic and Volk 2014). Its development and improvements consisted of the addition of tokenization, part-of-speech tagging, sentence segmentation, and sentence alignment. *CoStEP* was then used for aligning units at both the sentence and word level. It is worth mentioning that *Europarl* was not originally designed for linguistic purposes. One of the main obstacles for our project was the impossibility to distinguish original texts from translations. This gap was addressed in *CoStEP* by adding a function that gives access to important background information on speaker nationality, which is the best

[30] http://www.europarl.europa.eu/plenary/en/debates-video.html

proxy to differentiate between native and non-native speakers (i.e. between original and translated texts).

In a later stage of the project, a web-based application was developed for *CoStEP*, namely *Multilingwis,*[31] *the Multilingual Search Tool for Multi-word Units in Multiparallel Corpora* (Clematide, Graën and Volk 2016), which allows users, among other things, to find (complex) noun phrases and to explore translation variants across multiple languages. Throughout the project, this search tool was frequently used for preliminary investigations to test the data's potential before more systematic data retrieval processes.

4.1.3 Corpus alignment and annotation

In the first phase of the corpus improvement, the data provided by *Europarl* was subdivided into plenary sittings and classified by the date on which these events took place. Subsequently, the speaker turns were numbered in succession and separately for every language. Since it was not possible to rely on these numbers for accurate alignments, all other features available for individual speech turns were taken into consideration, i.e. the speaker's name, his/her affiliation, or the language used by the speaker for each single turn. Additionally, the tools and models provided by the TreeTagger (Schmid 1994) were used for tokenization, part-of-speech (PoS) tagging, and lemmatization of the corpus material. Tagging was computed with the language models available on the TreeTagger's website.[32] At that point, it was possible to segment speaker turns into sentences and, in a later stage, the statistical sentence aligner *hunalign* (Varga et al. 2005) was used for the sentence alignments. Additionally, the word alignment "was performed on the types of all tokens and on lemmas of content words" (Graën 2017: 10) with GIZA++ (see e.g. Gao and Vogel 2008) and the Berkeley Aligner (Liang et al. 2006). Finally, the corpus was parsed with the *MaltParser*[33] (Hall, Nilsson and Nivre 2006), which uses the dependency labels provided by the *Penn TreeBank II Annotation Scheme* (Bies et al. 1995).[34]

[31] https://pub.cl.uzh.ch/projects/sparcling/multilingwis/
[32] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/#Linux
[33] http://www.maltparser.org
[34] For a complete discussion regarding the corpus annotation and alignment, see Graën (2018).

Sections 4.2 and 4.3 are a methodology close-up and present two preliminary case studies. The former tests the reliability of the corpus for linguistic analysis and investigates whether the transcripts are faithful to the actual European Parliament speeches, whereas the latter analyses whether mixed texts (i.e. data comprised of both original texts and translations) can be used for the investigation of (variable) article use.

4.2 Transcripts vs. video recordings

The debates of the European Parliament are first recorded and then transcribed.[35] However, one disadvantage of the parliamentary transcripts is that they were not originally intended to be used for linguistic research. Monti et al. (2005: 119) state that the texts in the verbatim reports "do not reflect speech features very closely, as they undergo stylistic variation, punctuation is added and speakers' mistakes are amended (e.g. there are no instances of unfinished sentences, mispronounced words and ungrammatical structures)." Based on their assessment, the examination of the reliability of the corpus used for the present investigation is therefore essential. Transcript corpora are usually considered to be spoken data, but a comparison between the transcripts and the video recordings might call into question this assumption, as the linguistic accuracy of parliamentary transcriptions is not a primary goal for the transcribers.

The following is a case study, which examines the dissimilarities between the European Parliament debates' transcripts of three different days and their equivalent video recordings, which are freely available on the Internet.[36] In particular, it focuses on article use (definite article, indefinite article, and article omission) and takes into consideration two conditions which will be referred to as *groups* in the analysis: Group A regards the cases where the article is not transcribed but is produced by the

[35] According to the parliamentary Rule 194, verbatim reports have to respect the following points: 1. A verbatim report of the proceedings of each sitting shall be drawn up as a multilingual document in which all oral contributions appear in their original language. 2. Speakers may make corrections to typescripts of their oral contributions within five working days. […] 3. The multilingual verbatim report shall be published as an annex to the *Official Journal of the European Union* and preserved in the records of Parliament. 4. A translation into any official language of an extract from the verbatim report shall be made on request from a Member. […] (*Rules of Procedures of the European Parliament*, available at: http://www.europarl.europa.eu/sides/getLastRules.do?language=EN&reference=RULE-194&navigationBar=YES)

[36] Available at: http://www.europarl.europa.eu/plenary/en/debates-video.html

speaker during the parliamentary meeting, while Group B considers the cases in which the article is transcribed but is not used by the politician in the discussion. The analysis is mainly prompted by Mollin's (2007) study, whose findings show that the transcripts generally omit characteristics that are usually considered to belong to the spoken language, and that transcribers tend to modify the speakers' lexical and grammatical choices to make the texts more conservative. Furthermore, the two samples compared in Mollin's (2007) study belong to the same type of register and instance of communication. For this reason, the author believes that statistical tests are not needed, because the differences between the two samples cannot be either random or systematic. On the contrary, "each change is significant, because they are consciously introduced" (2007: 192). Thus, the decisions taken by the transcriber are crucial and well thought out. We will see that the spoken examples given in the analysis part (section 4.2.3) are not all grammatically correct according to English standard grammars, whereas other cases show variability that does not affect grammaticality. Does the transcriber make a change because the construction used by the speaker is ungrammatical, or because the variability is not well established in the written language? Surely, this is something that one has to consider while analyzing the differences between an original oral discourse and its equivalent transcribed speech. In the present case study, it is expected to find more differences between the oral discourse and the transcription in English rather than in German, because English articles are more likely to be omitted in the spoken medium (Biber et al. 1999: 267). Since most German nouns are accompanied by an article (Duden Online 2017), article use in German transcripts is expected to be more faithful to the original speeches.

Before proceeding to describe the data and show the results, it is important to discuss the differences between spoken, written language, and transcribed speech, and to briefly examine what previous research has argued on this topic.

4.2.1 Spoken vs. written language vs. transcribed speech

Investigating the relation between written and spoken language use is a continuing concern within linguistics and has been intensively studied by researchers. Koch and Oesterreicher (1985), for instance, propose their model of communication, which is based on the distinction between a graphic and phonic realization of the language on the one hand, and between *Sprache der Nähe* (i.e. language of immediacy) and

Sprache der Distanz (i.e. language of distance) on the other. Hence, language of immediacy and language of distance may be classified at opposite extremes. The universal features of the language of immediacy usually rely on the communication between two individuals, imply a face-to-face interaction, spontaneity, expressivity, and a reduction of morphosyntactic and lexical aspects; on the contrary, the universal features of the language of distance are generally established with the use of a text, are based on a spatiotemporal detachment, and imply reflection and planning. However, language of immediacy and language of distance are not strictly disconnected, as it might seem. On the contrary, Koch and Oesterreicher (1985: 17-18) claim that there is a continuum connecting them through different text types, such as *a* familiar conversation, *b* phone call with a friend, *c* interview, *d* printed interview, *e* diary entry, *f* personal letter, *g* job interview, *h* sermon, *i* talk, *j* broadsheet article, and *k* administrative regulations. Figure 4.1 is the simplified representation of this continuum.

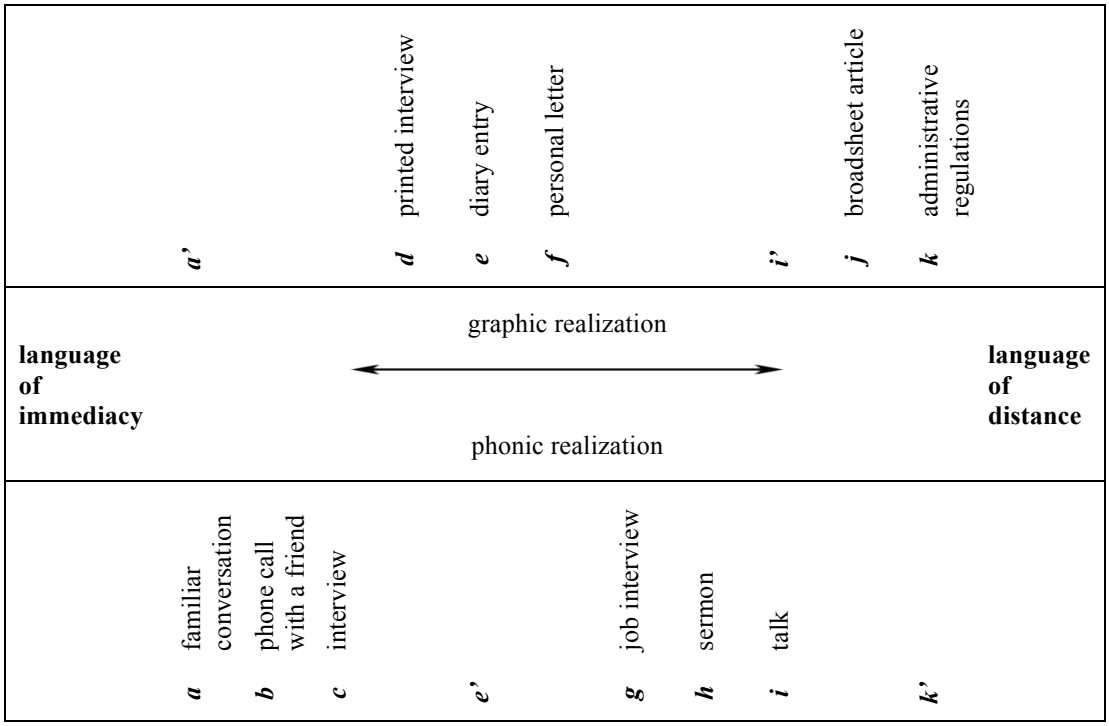

Figure 4.1: Representation of the continuum between language of immediacy and language of distance (adapted from Koch and Oesterreicher 1985: 18).

Language of immediacy and language of distance can be expressed in many variants and can always interchange some features. For instance, a phone call and a talk

conceptually belong to the phonic realization. But in comparison, a personal phone call is closer to the language of immediacy, while a talk is situated nearer to the pole of the language of distance. Similar examples can be found in the graphic realization, when comparing a diary (closer to the characteristics of the language of immediacy) and a broadsheet article (closer to the language of distance), for instance. Additionally, crossover patterns can also be observed, e.g. a birthday card, which is realized in the graphic medium but conceptually close to the language of immediacy, or a judge's speech in a courtroom, which is realized in the phonic medium but conceptually close to the language of distance. In other words, it is possible for many written types to find an equivalent representation in the language of immediacy and vice versa.[37] Therefore, the structure of their model is determined by the dichotomy between the graphic and phonic realization of language, and the continuum (or gradual features) between the realization in the language of immediacy and in the language of distance. According to this theory, even if some talks of the parliamentary members at the European Parliament might be prepared in advance, the transcripts could be considered as the transposition of the original speech into a written representation (i.e. transcriptions of relatively formal spoken language in an institutional context).

Halliday (1989: 30) also comments on the difference between spoken and written language and argues that writing mainly differs from speaking because it does not include all the various features of speech, like rhythm, intonation, degrees of loudness, variation in voice quality, pausing, and phrasing. Moreover, it is possible to have different forms of written language, and one of them is exactly the *transcribed speech*. Originally, writing down spoken discourse was done to preserve what was important for a family and its future generations. This then became culture, history, literature, and art (O'Connell and Kowal 1999: 104). Scholars have only recently had the chance to use electronic devices (e.g. the tape recorder) to make written transcriptions of natural speech. However, a transcribed text fails to represent the features of written language, as Halliday (1989: 41) states: "the main purpose of writing down speech, in fact, is to enable us to study speech, it is certainly not to provide a model of what written language ought to look like." Moreover, Green et al.

[37] Therefore, a diary can be read out loud (i.e. *e'*) and a talk can be printed or published (i.e. *i'*). The realizations *a'* and *k'* are considered extreme cases and mainly exist in one of the two realizations.

(1997: 172) affirm that "a transcript is a text that represents an event; it is not the event itself" and distinguish between transcriptions as an *interpretative process* and transcriptions as a *representational process*. The former is based on the choice made by the transcriber, and it is seen as a political act. In other words, the central point is *what* is transcribed. The latter is defined as a partial representation, in which the ways data are represented influence the possible meanings and interpretation. Thus, the focus is on *how* it is transcribed. On the other hand, Bucholtz (2000) classifies transcriptions into two different types: *naturalized* and *denaturalized*. The process of naturalized transcriptions is "less visible through literacization" (Bucholtz 2000: 1461), and a written style of the language is preferred over oral discourse features. On the contrary, denaturalized transcriptions are more faithful to the speech and are therefore harder to interpret for the reader. It is thus important to underline that, when analyzing transcriptions, the goals of the transcriber are the first and most relevant elements that need to be taken into account. As pointed out by Bucholtz (2000: 1463), "transcription is an act of interpretation and representation, it is also an act of power" and, therefore, we need to acknowledge this problem.

Previous research on this topic includes Walker (1990), Coulthard (1996), and Bucholtz (2000) on the discrepancies of transcripts in courtroom proceedings and police interrogations, and Slembrouck (1992) and Mollin (2007) on the comparison of officially released transcripts of Parliamentary debates to a linguistically accurate transcription of the original spoken debates. All studies reached the same conclusion, stating that the original speeches greatly differ from the officially released transcripts. In particular, Slembrouck (1992) and Mollin (2007), who focus on the minutes of the proceedings of the British Parliament and use the Hansard data as a corpus, affirm that there are considerable discrepancies between the two types of transcripts they investigated. Firstly, some elements from the original contribution are omitted in the transcript (e.g. repetitions, false starts, and reformulations). Secondly, other elements are deleted (e.g. incomplete utterances, mistakes, redundancies, and pauses). Slembrouck (1992: 104) describes this effect as a "filtering out of disfluency and other obvious properties of spokenness." Transcripts of parliamentary debates are thus different from what we usually refer to as transcribed speech. As stated by Tannen (1984: 245), "[i]n general, transcripts do not feel to readers like 'real' conversation – they are not immediately intelligible like the dialogue in a novel or a movie, they don't get to a point, they don't really begin or end."

The turn provided in Table 4.1 is a good example of a contrast between a politician's original speech and its transcript available from the website of the European Parliament.[38] This speaker turn[39] enables us to note some features of spoken language that were present (or absent) in the politician's original spoken delivery of the speech but modified in its equivalent transcript, such as: the deletion of filler words (e.g. *ehm*), repetitions (e.g. *a* and *charges*) and self corrections (e.g. *disrep-pute* and *disha-disharmony*), the addition of title nouns (e.g. *Mr*), the use of non-contracted forms (e.g. *he had been detained*), and the exclusion of final greetings or thanks (e.g. *Thank you*).

Original speech	**Transcript**
<*ehm* i wanted to raise with you *president* quite *a a* serious case about a sri lankan journalist j. s. tissainayagam which we raised on our recent delegation visit to sri lanka>	**Jean Lambert (Verts/ALE). -** Mr President, I wanted to raise with you quite a serious case about a Sri Lankan journalist, Mr J. S. Tissainayagam, which we raised on our recent delegation visit to Sri Lanka.
<he is a very well known writer and journalist and has been running amongst other things a german government funded website called outreach promoting peace and justice>	He is a very well-known writer and journalist and has been running, amongst other things, a German-Government-funded website called 'Outreach' promoting peace and justice.
<at that point *he'd been detained* without charge for over four months in poor conditions and he was finally charged and remanded in custody last week under the country's prevention of terrorism *charges* with *charges* related to bringing the government into *disrep-pute* and stirring up communal *disha-disharmony*>	At that point he had been detained without charge for over four months in poor conditions, and he was finally charged and remanded in custody last week under the country's Prevention of Terrorism Act, with charges related to bringing the government into disrepute and stirring up communal disharmony.
<we would ask you *president* to use your good offices with council and commission to follow this important case not least to see to it that he be able to meet his lawyers in private which he *hasn't been able to do* yet and that there will be full disclosure at the evidence against him>	We would ask you, Mr President, to use your good offices with Council and Commission to follow this important case, not least to see to it that he be able to meet his lawyers in private – which he has not been able to do yet – and that there will be full disclosure of the evidence against him.
<*thank you*>	----

Table 4.1: Comparison between the original speech and its transcript of a speaker turn.

[38] Note that the equivalent text from the *Europarl* corpus entirely matches the transcript available on the website of the European Parliament.
[39] Source: *Europarl*, 2008-09-01.xml

More examples of discrepancies between the video recordings and the official transcripts from the European Parliament website are listed in Table 4.2.[40]

Variation type	Video recording (example)	Transcript
Word order	a. vorab möchte ich allen meine gratulation aussprechen zur erreichten einigung	a'. Vorab möchte ich allen meine Gratulation zur erreichten Einigung aussprechen!
	b. ho sostenuto nel mio ruolo di relatore	b'. […] nel mio ruolo di relatore ho sostenuto […].
	c. unfortunately what we see and what we have on the table here	c'. Unfortunately, what we see and what we have here on the table […].
Deletion or modification of words	d. hier muss ich sagen wenn die kommission die hüterin der verträge ist so ist das europäische parlament offensichtlich die hüterin der demokratie	d'. Hier muss ich sagen, wenn die Kommission die Hüterin der Verträge ist, so ist das Europäische Parlament offensichtlich der Hüter der Demokratie.
	e. i recall clearly as rapporteur dealing with the eu india free trade agreement	e'. I recall clearly, as the rapporteur dealing with the EU-India Free Trade Agreement […].
	f. it is vital and we have got to respond to the situation	f'. It is vital and we have to respond to the situation.
Deletion or modification of informal expressions/words	g. wir haben uns dann immer wieder	g'. Wir haben uns immer wieder […].
	h. the european patent office doesn't care about this	h'. The European Patent Office does not care about this.
	i. so i urge all colleagues please to support	i'. I urge all colleagues, please, to support […].
Addition of words	j. our response 40 years in the waiting	j'. Our response was 40 years in the waiting.
Substitution of foreign words	k. research	k'. Forschung
	l. agreement	l'. Übereinkommen
Use of indirect speech instead of direct speech	m. terminaré citando al prestigioso instituto europeo sobre propiedad intelectual y derecho a la competencia max planck que dice lo siguiente la patente unitaria	m'. Terminaré citando al prestigioso Instituto Europeo Max Planck para la Propiedad Intelectual y el Derecho de la Competencia, que señala que la patente unitaria […].

Table 4.2: List of variation types of discrepancies between video recordings and parliamentary transcripts.

[40] Note that the examples included in Table 4.2 are taken from the transcripts analysed in the current case study.

Changes concern word order, deletion or modification of words and informal expressions, addition of words, substitution of foreign words, and use of indirect instead of direct speech. Particularly interesting is example *e* and its equivalent *e'*. In the original speech, the speaker does not use an article in front of *rapporteur*, but the transcriber adds the definite article in the transcription. This example is relevant for the present study, because it underlines a case of article variability in English. In short, both constructions are acceptable, but the transcriber decided to use a more conservative form using the definite article.

In conclusion, Table 4.2 shows that parliamentary transcripts do not perfectly correspond to the original politicians' speeches. Therefore, one has to remember that researchers always need to keep in mind that "transcriptions made for purposes other than linguistic ones need to be assessed for their linguistic accuracy and thus reliability if chosen for linguistic study" (Mollin 2007: 208).

4.2.2 Preliminary case study: data description

The debates chosen for this analysis are taken from 11 December 2012, 1 July 2013, and 21 October 2014. The criteria for selecting the debates were as follows: firstly, the equivalent video of the transcript had to be available;[41] secondly, the transcript had to be downloadable in *pdf* format;[42] and thirdly, the turns had to be in the original language used by the speaker, in other words, no translations were included in the analysis.

The current study excludes the information that does not represent transcribed speech (e.g. the main points of the debates or the general information on the meeting, such as the exact time of the beginning of the session) and only analyzes the turns transcribed in English and German. Unfortunately, it is common that speakers do not use their mother tongue to communicate. Therefore, among the turns transcribed in English or in German, the ones that were produced by members of Parliament whose

[41] Debates have been recorded and transcribed since July 1999. Thus, both videos and transcripts are freely available and are published in their original language (*European Parliament plenary debates*, available at: http://www.europarl.europa.eu/plenary/en/debates-video.html).

[42] Note that the transcripts can be downloaded and saved as pdf files only starting from June 2007.

mother tongue was not English or German were not considered.[43] Table 4.3 and Table 4.4 exhibit the distribution of speakers that use English and German respectively, in the three debates. Both tables also give the number of analyzed turns. All in all, this sample contains 123 members of Parliament who use English to communicate, but, as the data show, 46% of these speakers are not English native speakers. The situation is different for German. Most speakers, i.e. 87 out of 93, have German as their first language.

ENG				
	11 Dec 12	**1 Jul 13**	**21 Oct 14**	**Total**
Speakers talking in English	51	22	50	**123**
Native speakers	30 (59%)	16 (73%)	21 (42%)	**67 (54%)**
Non-native speakers	21 (41%)	6 (27%)	29 (58%)	**56 (46%)**
Analyzed turns (only native speakers)	58 (12.400 words)	20 (5.103 words)	29 (5.277 words)	**123 (22.780 words)**

Table 4.3: Distribution of speakers talking in English and number of turns analysed for the investigation.

GER				
	11 Dec 12	**1 Jul 13**	**21 Oct 14**	**Total**
Speakers talking in German	39	20	34	**93**
Native speakers	38 (97%)	17 (85%)	32 (94%)	**87 (93%)**
Non-native speakers	1 (3%)	3 (15%)	2 (6%)	**6 (7%)**
Analyzed turns (only native speakers)	56 (12.410 words)	36 (6.547 words)	45 (9.452 words)	**137 (30.409 words)**

Table 4.4: Distribution of speakers talking in German and number of turns analysed for the investigation.

Moreover, English and German are the first official languages in multiple European countries. For this reason, it is interesting to examine the different speakers' nationalities. Table 4.5 refers to English and shows that 58 speakers are originally from Britain and only 10 from Ireland. Thus, 47% of the speakers are British, 8% are

[43] Note that speakers' nationalities are based on the politician's personal information, provided on the website of the European Parliament, available at: http://www.europarl.europa.eu/meps/en/full-list.html

Irish, and the remaining 45% come from various European countries[44], whose first official language is not English.

Nationalities – ENG				
	11 Dec 12	**1 Jul 13**	**21 Oct 14**	**Total**
British	26 (51%)	14 (64%)	18 (36%)	58 (47%)
Irish	4 (8%)	2 (9%)	3 (6%)	10 (8%)
Other	21 (41%)	6 (27%)	29 (58%)	56 (45%)

Table 4.5: Distribution of nationalities of the speakers talking in English.

Nationalities – GER				
	11 Dec 12	**1 Jul 13**	**21 Oct 14**	**Total**
German	27 (69%)	11 (55%)	25 (74%)	63 (68%)
Austrian	11 (28%)	6 (30%)	7 (20%)	24 (26%)
Other	1 (3%)	3 (15%)	2 (6%)	6 (6%)

Table 4.6: Distribution of nationalities of the speakers talking in German.

Table 4.6 shows that in the German sub-sample, the majority come from Germany (63 members), and that only 24 speakers come from Austria. This means that 68% are German, 26% are Austrian, and the remaining 6% come from a different European country, whose first official language is not German.[45]

4.2.3 Preliminary case study: results and analysis

As previously mentioned, two groups are taken into account for the investigation: the first one (Group A) and concerns the cases where there is no article in the transcription, but it is produced by the member of Parliament in the original speech, whereas the second one (Group B) refers to the cases in which an article is inserted into the transcription but is not used by the speaker in the delivery of his/her speech. The first step in this process was to annotate the data. The annotation was done manually, comparing the transcripts with the video recordings of the meetings. Thus, the methodology consisted of reading the verbatim report while listening to the audio

[44] English non-native speakers come from the following countries: Belgium, Bulgaria, Czech Republic, Estonia, Germany, Italy, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Romania, and Sweden.

[45] German non-native speakers come from the following counties: Belgium, France, and Slovakia.

of the equivalent videos and marking every single case of interest directly on the *pdf* file. As stated before, both definite and indefinite articles were considered for the study, and both singular and plural forms were included in the analysis. Cases where the article was substituted by the transcriber with another determiner, like a demonstrative or a possessive, were not taken into consideration. An example of these excluded cases is given in (1)a and (1)b.[46]

(1)

(a) Vor ein paar Wochen bin ich in meinem Wahlkreis in einem kleinen, hochinnovativen Technologieunternehmen gewesen, wo mir im Rahmen ihrer Präsentation auch vorgestellt wurde, [...] Das sind die drei roten Linien, wo wir gesagt haben, wenn wir überhaupt auf dem Weg mitmachen, müssen sie respektieren werden. *(12-11-2012, German, VOICE)*

(b) Vor ein paar Wochen bin ich in meinem Wahlkreis in einem kleinen, hochinnovativen Technologieunternehmen gewesen, wo mir im Rahmen einer Präsentation auch vorgestellt wurde, [...] Das sind die drei roten Linien, wo wir gesagt haben, wenn wir überhaupt auf diesem Weg mitmachen, müssen sie respektieren werden. *(12-11-2012, German, CORPUS)*

In (1)a, the speaker first uses a possessive pronoun and then a definite article. In (1)b, the transcriber substitutes them with an indefinite article and a demonstrative, respectively. In this example, the determiner slot is always filled, but with a different determiner. Since the present case study focuses on the addition or omission of articles in the transcripts, these cases are not part of the investigation.

After annotation, all cases were analyzed to determine whether there was a difference concerning the presence or the omission of articles. Table 4.7 gives the exact number of relevant cases, found in the verbatim reports of the proceedings of the three sittings. The results show that English displays more differences than German in both groups: in 37 cases an article is present in the transcript but is not used by the speaker, and in 38 cases an article is pronounced by the speaker but is not transcribed in the report. By contrast, in German the number of discrepancies between original and transcript is evidently lower than in English. This might mean that German transcribers tend to be closer to politicians' speeches. More specifically, the findings show 19 cases where the article is present in the transcript but is not uttered

[46] Throughout the chapter, the first example is always the original spoken example (marked with *VOICE*), the second one the Parliament transcript (marked with *CORPUS*).

by the speaker, and only 9 cases where the article is present in the video recording but is omitted in the transcript.

	11 Dec 12		1 Jul 13		21 Oct 14		**Total**	
	ENG	**GER**	**ENG**	**GER**	**ENG**	**GER**	**ENG**	**GER**
Group A	25	4	8	4	4	1	**37**	**9**
Group B	21	14	13	3	4	2	**38**	**18**
Total	46	18	21	7	8	3	**75 (4%)**	**28 (1%)**

Table 4.7: Number of cases found in the three transcripts in Group A and Group B, regarding English and German.

The following is a detailed qualitative analysis of the results. The first subsection includes the classification of nouns that follow the added or omitted article in each group. The different categories used in the analysis are the following: abbreviation (e.g. *NATO*, *UN*), abstract noun (e.g. *justice*, *freedom*), non-abstract noun (e.g. *truck*, *agenda*), human (e.g. *citizen*, *refugee*), proper noun (e.g. *Mr. Ratas*, *France*), institution (e.g. *Parliament*, *the Department of Trade*), and temporal expression (e.g. *at the end*). The second subsection investigates the number of the nouns in both groups, i.e. whether they are used in the singular or plural form. Finally, the examination aims to study how many of the relevant cases are related to the definite article, and how many are related to the indefinite article. This will help to see whether the addition or omission of articles affects more nouns that are perceived as specific, i.e. used with the definite article, or non-specific, i.e. used with the indefinite article.

A. NOUN CLASSIFICATION

Figure 4.2 exhibits the distribution of the nouns in English and German among the different noun categories in Group A.

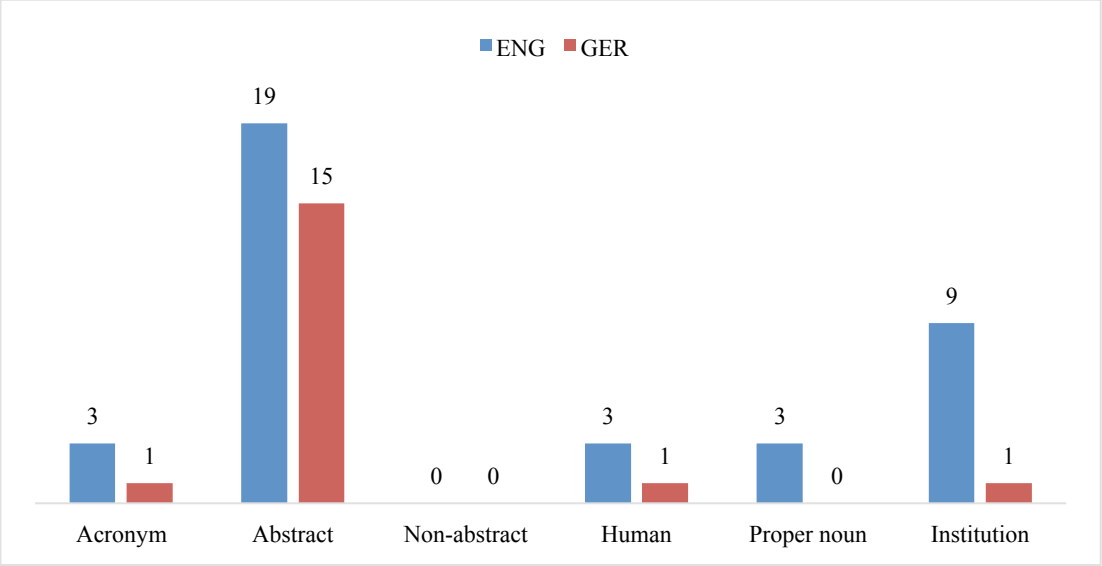

Figure 4.2: Comparison between English and German of the noun types distribution in Group A (raw numbers).

The first category concerns the nouns classified as abbreviations, as the examples (2)a, (2)b, and (3)a, (3)b show.

(2)

(a) Our experience of multihandling in Ø UK has been successful for many years. *(12-11-2012, British, VOICE)*

(b) Our experience of multihandling in the UK has been successful for many years. *(12-11-2012, British, CORPUS)*

(3)

(a) [...] das heißt, er möchte, dass Ø EuGH aus dem ganzen Verfahren herausgenommen wird, der EuGH soll [...]. *(12-11-2012, German, VOICE)*

(b) [...] das heißt, er möchte, dass der EuGH aus dem ganzen Verfahren herausgenommen wird, der EuGH soll [...]. *(12-11-2012, German, CORPUS)*

Figure 4.2 shows that in English 3 cases are abbreviations, while in German there is only one. (3)a is particularly interesting because the speaker mentions the same abbreviation twice: he omits the article in front of the first abbreviation but uses it in the second one.

The second category relates to abstract nouns. The English sub-sample has 19 cases, while the German one has 15. (4)a and (4)b, and (5)a, (5)b are relevant examples.

(4)

(a) […] big public expenditure and large government deficits do not seem to be providing Ø solution. *(12-11-2012, British, VOICE)*
(b) […] big public expenditure and large government deficits do not seem to be providing the solution. *(12-11-2012, British, CORPUS)*

(5)

(a) Ø Menschenrechte hat mein Vorredner angeschnitten. *(12-11-2012, Austrian, VOICE)*
(b) Die Menschenrechte hat mein Vorredner angeschnitten. *(12-11-2012, Austrian, CORPUS)*

The construction used by the speaker in (4)a is an example of ungrammaticality, which is corrected by the transcriber in (4)b. On the other hand, in (5)a and (5)b we have a case of variability in German. The definite article in front of the plural noun *Menschenrechte* can be omitted without making the sentence ungrammatical. In other words, the speaker uses a generic reference, while the transcriber prefers specificity.

The non-abstract nouns category has no results, neither in English, nor in German. On the other hand, the category referring to human beings presents 3 cases in English and only 1 in German. It is worth noting that in the example provided in (6)a and (6)b, the noun *rapporteur* can also be considered a unique role, task, or professional position, in such cases article omission can alternate with *the* (Quirk et al. 1985: 276).

(6)

(a) I recall clearly, as Ø rapporteur dealing with the EU-India Free Trade Agreement negotiations. *(12-11-2012, British, VOICE)*
(b) I recall clearly, as the rapporteur dealing with the EU-India Free Trade Agreement negotiations. *(12-11-2012, British, CORPUS)*

(7)

(a) Das Problem ist nur, dass es Kollegen gibt, die sich zu Wort melden, wie Ø Kollege Mölzer [...]. *(12-11-2012, Austrian, VOICE)*
(b) Das Problem ist nur, dass es Kollegen gibt, die sich zu Wort melden, wie der Kollege Mölzer [...]. *(12-11-2012, Austrian, CORPUS)*

The proper nouns category yields 3 cases in English and no cases in German. In the following example, (8)a and (8)b, both are country names:

(8)

(a) […] yet the following EU countries have not yet ratified it: Austria, Belgium, Ø Czech Republic, Estonia, Germany, Hungary, Ireland, Italy, Lithuania, Portugal, Romania, Slovakia, Slovenia, Ø United Kingdom and, our newest member, Croatia. *(07-01-2013, British, VOICE)*

(b) […] yet the following EU countries have not yet ratified it: Austria, Belgium, the Czech Republic, Estonia, Germany, Hungary, Ireland, Italy, Lithuania, Portugal, Romania, Slovakia, Slovenia, the United Kingdom and, our newest member, Croatia. *(07-01-2013, British, CORPUS)*

In (8)a, the speaker omits the article in front of *Czech Republic* and *United Kingdom*. In this case, the absence of the definite article may be determined by the fact that these two place names are included in a list, but this omission is considered ungrammatical by the transcriber, who inserts the article in the transcription.

(9)a, (9)b and (10)a, (10)b are examples of the final noun category, namely institutions.

(9)

(a) […] I would, at this point, declare an interest as vice-president of the Scottish Society for Ø Prevention of Cruelty to Animals. *(12-11-2012, British, VOICE)*

(b) […] I would, at this point, declare an interest as vice-president of the Scottish Society for the Prevention of Cruelty to Animals. *(12-11-2012, British, CORPUS)*

(10)

(a) [...] das heißt natürlich auch, dass die Rechte Ø Europäischen Parlaments als Gesetzgeber beeinträchtigt werden. *(12-11-2012, German, VOICE)*

(b) [...] das heißt natürlich auch, dass die Rechte des Europäischen Parlaments als Gesetzgeber beeinträchtigt werden. *(12-11-2012, German, CORPUS)*

English includes 9 instances with institutional nouns. It is interesting to note that in (9)a and (9)b, the institutional noun *Prevention of Cruelty to Animals* is only a part of a more complex multi-word institutional noun. The use of the definite article in the transcription is explained by the fact that the official name of this institution contains the article in front of *Prevention*. The use of the definite article, however, is also proved by the study conducted by Tse (2003), who shows that the strongest grammatical factor triggering definite article use with multi-word institutional nouns is the prepositional phrase used as postmodifier, the preposition *of* in our case. On the other hand, German has only one case where the speaker uses the suffix *–s* after the noun, which represents the genitive case, even if he omits the definite article in front of it. This omission is ungrammatical; therefore, the change in the transcription is

towards grammatically correct standard German usage (i.e. this instance is not a case of article variability).

As mentioned before, Group B focuses on the cases where the speaker produces an article during the meeting, but it is omitted in the transcript. The results are very similar. The examples are also distributed among the different categories of nouns that follow the relevant cases. Figure 4.3 shows the results of the analysis.

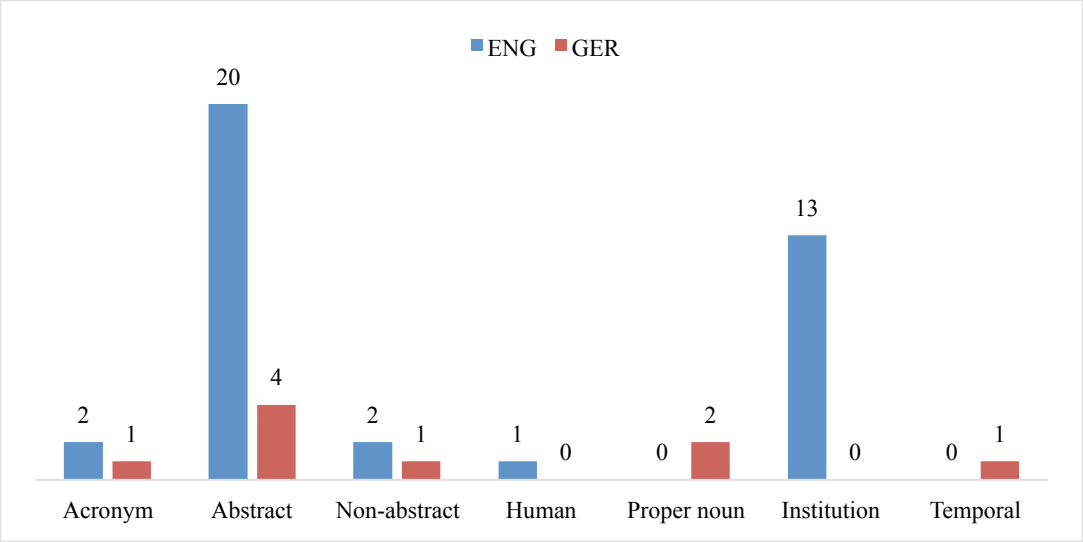

Figure 4.3: Comparison between English and German of the noun types distribution in Group B (raw numbers).

(11)a and (11)b are one of the two cases found in English in the abbreviation category, (12)a and (12)b refer to the single case that appeared in German:

(11)
- (a) […] quite clearly it is important that we encourage the SMEs across the EU […]. *(12-11-2012, British, VOICE)*
- (b) […] quite clearly it is important that we encourage SMEs across the EU […]. *(12-11-2012, British, CORPUS)*

(12)
- (a) Und das ist für mich ein ganz entscheidendes Beispiel dafür, warum das, was wir heute beschließen, für die KMU ein gewaltiger Schritt nach vorne ist. *(German, VOICE)*
- (b) Und das ist für mich ein ganz entscheidendes Beispiel dafür, warum das, was wir heute beschließen, für Ø KMU ein gewaltiger Schritt nach vorne ist. *(German, CORPUS)*

The abstract nouns category has more examples in English, 20 cases, and fewer in German, only 4. (13)a, (13)b and (14)a, (14)b are examples of this classification.

(13)

(a) […] the importance of promoting technological developments for the education and the protection of minors. *(12-11-2012, British, VOICE)*
(b) […] the importance of promoting technological developments for the education and Ø protection of minors. *(12-11-2012, British, CORPUS)*

(14)

(a) Zuerst möchte ich unserem Berichterstatter einen herzlichen Dank aussprechen [...]. *(12-11-2012, German, VOICE)*
(b) Zuerst möchte ich Ø herzlichen Dank aussprechen [...]. *(12-11-2012, German, CORPUS)*

In (13)b, it is possible that the transcriber omits the article, because his or her intention is to be closer to the written standard language and s/he prefers to avoid repetition: *the* is first used in front of the word *education*, and it is therefore reasonable that the transcriber does not want to repeat it a second time.

As already seen before, the category of non-abstract nouns is not amply represented in the dataset. English only has 2 cases, (15)a and (15)b is one of them, and German only one. The following is a case found in English:

(15)

(a) […] there is also the role of the goods and the materials which are transferred as well as knowledge. *(12-11-2012, British, VOICE)*
(b) […] there is also the role of the goods and Ø materials which are transferred as well as knowledge. *(12-11-2012, British, CORPUS)*

In (15)b, as in (13)b, the transcriber omits the article, because it is previously used in front of the word *goods*; hence, it is highly possible that the transcriber, again, simply prefers to avoid repetition.

The category of nouns regarding humans is not large either: (16)a and (16)b is the exceptional example in English.

(16)

(a) Let us look at the UK's Intercept Programme which is going to collect all the data of the UK citizens […]. *(12-11-2012, British, VOICE)*
(b) Let us look at the UK's Intercept Programme which is going to collect all the data of Ø UK citizens […]. *(12-11-2012, British, CORPUS)*

The proper noun category has no cases in English and 2 cases in German, (17)a, (17)b is one of the two examples.

(17)

(a) Die Frau Kollegin Gál hat völlig Recht [...]. *(Austrian, VOICE)*
(b) Ø Frau Kollegin Gál hat völlig Recht [...]. *(Austrian, CORPUS)*

The transcriber omits the article in (17)b because the use of the definite article in German in front of proper nouns is usually informal (Rowlinson 1994: 90). The speaker comes from Austria, and the use of an article in this context might also be related to an influence of local dialects.

The category referring to institutional nouns is definitely richer in English; there are, in fact, 13 cases. By contrast, no cases occur in German. The following is an English example.

(18)

(a) […] when he was responding to the result of the vote here in *the Parliament*. […] As he pointed out, we now have three-way agreement: the Commission, the Council and the Parliament. *(12-11-2012, Irish, VOICE)*
(b) […] when he was responding to the result of the vote here in *Parliament*. […] As he pointed out, we now have three-way agreement: the Commission, the Council and Ø Parliament. *(12-11-2012, Irish, CORPUS)*

It has to be noted that the case above comes from an Irish speaker. As already explained in section 2.8, IrE strongly differs from BrE and tends to use an article more often than any other standard forms of English (see for instance Hickey 2007, Corrigan 2010, Kallen 2013).

The last category deals with temporal expressions. It only includes one exceptional case of the German sub-sample.

(19)

(a) [...] ich erwarte, dass wir uns am Ende des nächsten Haushaltsjahres auf einige Schwierigkeiten zubewegen werden [...]. *(12-11-2012, German, VOICE)*
(b) [...] ich erwarte, dass wir uns Ø Ende des nächsten Haushaltsjahres auf einige Schwierigkeiten zubewegen werden [...]. *(12-11-2012, German, CORPUS)*

The speaker uses the definite article combined with the preposition *an* in front of *Ende*. On the other hand, in the transcription a bare noun is used instead. This is a case of variability, because both cases are grammatically correct. However, the use of the article in this context is more colloquial. Hence, this case further proves that

transcribers have a stronger tendency to adopt a more formal style, closer to the written language.

B. NUMBER AND ARTICLE DISTRIBUTION

Examining whether the changes in the transcripts occur more with singular or plural nouns is interesting because plural countable nouns, in particular, can express the generic reference in different ways, with either the definite article or the zero article (Biber et al. 1999: 265). Moreover, Biber et al. (1999: 267) state that the use of the indefinite article is relatively similar in both written and spoken language; on the contrary, the definite article shows more differences. Since all these aspects can provide cases of variability, it is worth investigating the number of each NP and the article distribution of the changes. This will help to indicate whether the discrepancies between the parliamentary transcripts and the video recordings occur more with singular or plural nouns, and whether they lean toward the definite or the indefinite article.

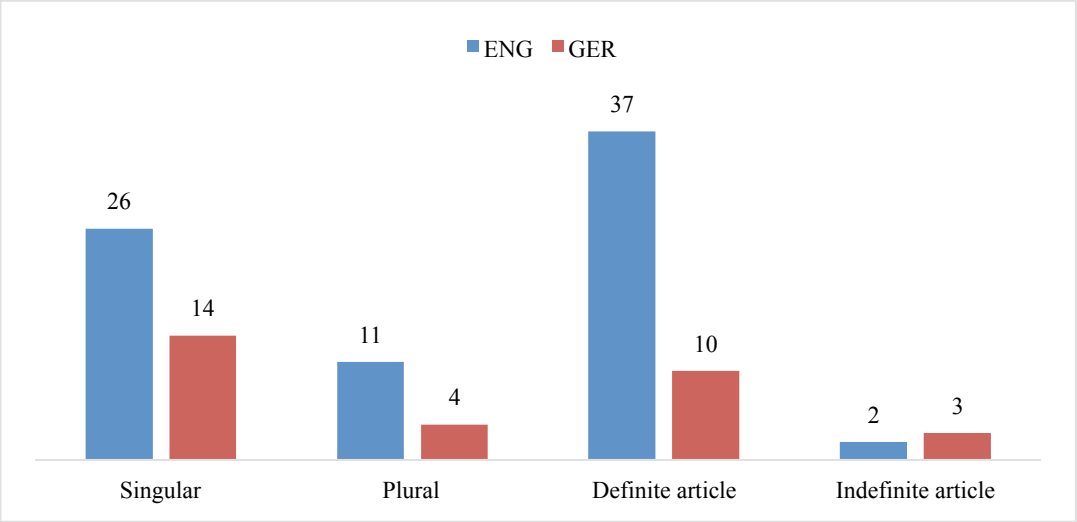

Figure 4.4: Comparison between English and German on noun number and article distribution in Group A (raw numbers).

Figure 4.4 refers to Group A and puts together the number and article distribution. The results show that in English, 26 cases are singular nouns and 11 are plural nouns. By contrast, in German 14 cases are singular and 4 are plural. The distribution of the definite and indefinite article is similar in both languages. Specifically, the results

exhibit a higher use of the definite article. In English 37 cases occur with the definite and only 2 cases with the indefinite article, while in German there are 10 cases with the definite and 3 with the indefinite article. Figure 4.5 shows that the results of Group B are similar to the findings of Figure 4.4. English has 23 cases in singular and 15 in plural. On the other hand, German only has one plural case and 8 singular cases. The distinction between the usage of the definite and indefinite article is also very similar to the previous group: whereas in English the definite article occurs in 35 cases and the indefinite only in 3, in German the definite article occurs in 8 cases and the indefinite only one time.

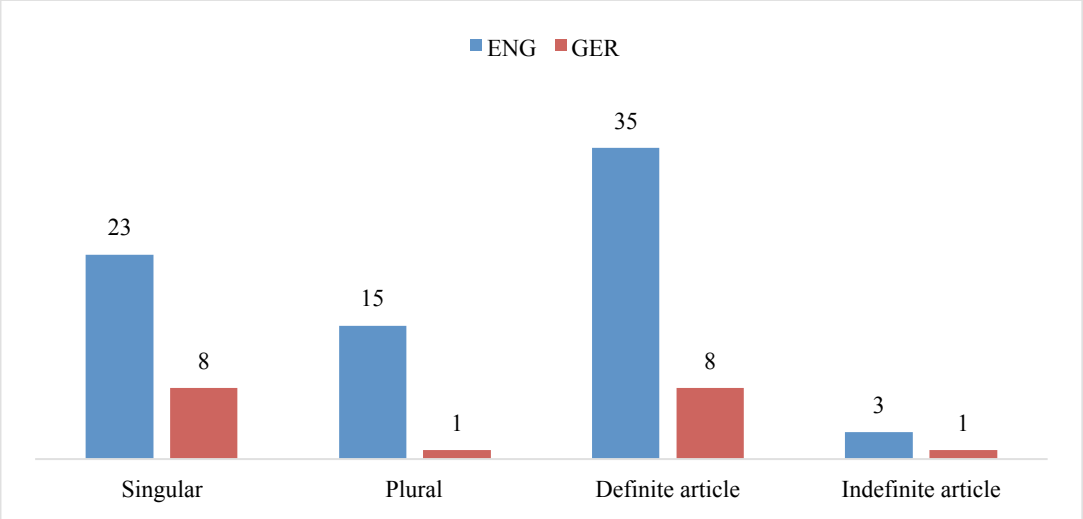

Figure 4.5: Comparison between English and German on noun number and article distribution in Group B (raw numbers).

The changes found in the transcriptions mainly occur with singular nouns. However, the analysis in the previous section has shown that in both groups the majority of cases belong to the category of abstract nouns, which cannot always be pluralized. The results of the article distribution show that most changes occur with the definite article rather than the indefinite article. This supports the previously mentioned observation proposed by Biber et al. (1999), which states that the definite article exhibits more differences between written and spoken language than the indefinite article. At this point, it is relevant to examine the nature of the changes. In other words, did the transcriber add or omit an article because of grammaticality, or because of variability in language? The following subsection will try to answer this question.

C. VARIABILITY VS. GRAMMATICALITY

Every change found in a transcript represents an intended decision, which was taken by the transcriber. A key aspect that one has to consider while analysing transcribed speech is whether the changes are due to variability or grammaticality. Put differently, it is important to know whether the transcriber makes a change because the construction used by the speaker is not grammatically correct according to the standard grammar, or because the construction used in the original speech is variable and not yet established in the written language.

The pie charts below provide an overview of all changes and compare English and German in the two different groups included in the analysis. Group A refers to the cases where the article was originally produced by the speaker but was omitted in the transcript. Figure 4.6 shows that the results in English and in German have the same tendencies. In both languages, more than half of the changes are due to variability. However, the frequency is higher in English than in German, with percentages at 89% and 55%, respectively. On the other hand, Group B refers to those cases where the article was not used by the speaker but was added by the transcriber. The results, shown in Figure 4.7, are very similar in both languages, whose changes are mainly related to variability (almost 55%).[47]

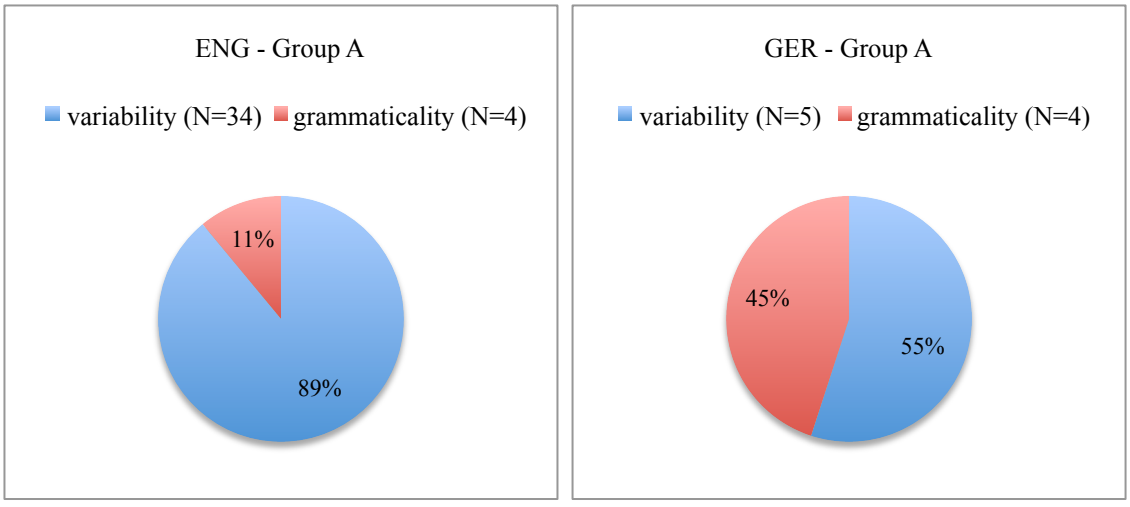

Figure 4.6: Comparison between English and German related to variability and grammaticality in Group A.

[47] Note that most of the grammatically incorrect cases in German come from the same speaker, who is then exceptional.

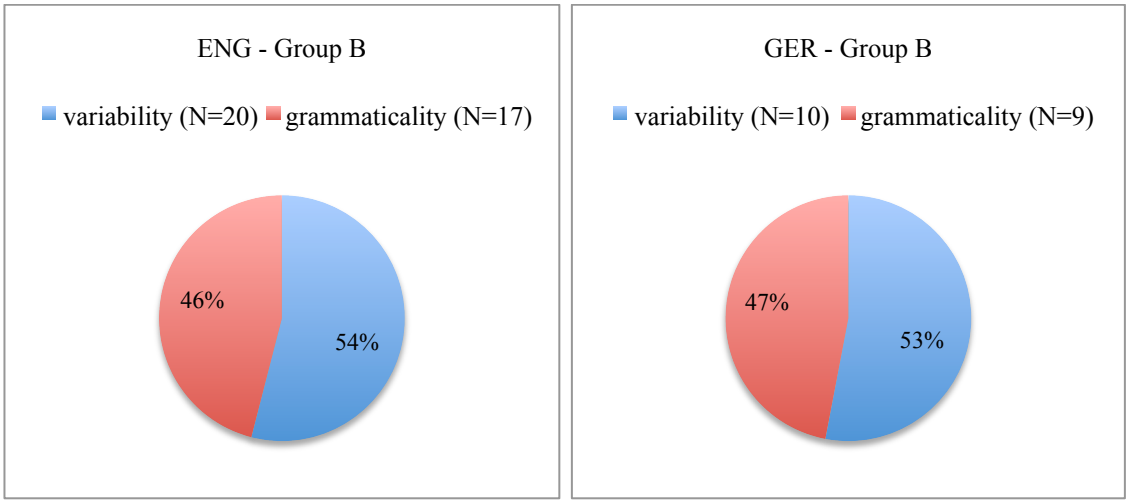

Figure 4.7: Comparison between English and German related to variability and grammaticality in Group B.

It is worth pointing out that there can be different categories of grammaticality. Firstly, transcribers may face unclear cases. For instance, the speaker may pronounce the article in an incomprehensible way. In other words, phonetic reasons can be the cause of the change shown in (20)a and (20)b. Moreover, a case that might seem like a grammatical mistake, can turn out to be an ongoing change in the spoken medium. An example is (21)a and (21)b, where the transcriber decided to add the definite article, because it is required by standard grammars. However, this construction is not considered grammatically incorrect in spoken language.

(20)
(a) […] when I was studying the digital world, one of Ø things that I found interesting was that […]. *(12-11-2012, S. Kamall, British, VOICE)*
(b) […] when I was studying the digital world, one of the things that I found interesting was that […]. *(12-11-2012, S. Kamall, British, CORPUS)*

(21)
(a) […] before we resort to demanding more from Ø taxpayer. *(12-11-2012, R. Ashworth, British, VOICE)*
(b) […] before we resort to demanding more from the taxpayer. *(12-11-2012, R. Ashworth, British, CORPUS)*

Another category regards cases in which the variable is present in the spoken medium only and cannot be accepted in written language, as shown in the following example.

(22)
(a) […] we got the Nobel Peace Prize, Ø wonderful achievement for us. (*12-11-2012, S. Kelly, British, VOICE*)

(b) […] we got the Nobel Peace Prize, which was a wonderful achievement for us. (*12-11-2012, S. Kelly, British, CORPUS*)

In spoken language, the construction without an article in (22)a is not grammatically incorrect but cannot be used in written language. This proves that the spoken medium is more tolerant of ellipsis.

Overall, the results show that in both English and German, the majority of changes are due to variability. The transcriber therefore makes changes because the construction used in the spoken language is not well established in the written medium. In other words, they generally move towards a more conservative style. The transcriber cleans the speech from those constructions that are not yet accepted by standard grammars, making the transcribed speech closer to the written language.

4.3 Original texts vs. translations

One of the main goals of the European Parliament and its institutions is to reinforce the model of a multicultural society and to preserve the unique linguistic diversity of Europe. Every parliamentary member has the right to speak in an official language of his or her choice. Moreover, since many citizens speak only one language, the European Parliament is determined to provide them access to legislation, procedures and information in their mother tongue. Its translation services are considered one of the largest in the world. According to the European Parliament website, over 2,000 translators and 800 interpreters are required every day and since 2005 over a million pages are translated each year (*European Parliament – never lost in translation*, OD[48]).

One of the biggest methodological obstacles one has to face when working with *Europarl* is related to the distinction between original texts (i.e. transcripts of speakers' productions) and their corresponding translations. Data annotated with the names and (language) background of the parliamentary members would be helpful to make this distinction. Unfortunately, at the initial stage, the corpus used for this

[48] Online Document, available at: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+IM-PRESS+20071017FCS11816+0+DOC+XML+V0//EN. For instance, the staff of the European Commission, one of the European Parliament's institutions, is composed of 1,750 linguists and 600 advocates, supported by 600 full-time and 3,000 freelance interpreters (European Commission, Supporting language learning and linguistic diversity, OD available at: http://ec.europa.eu/languages/policy/linguistic-diversity/official-languages-eu_en.htm).

project was not consistently tagged with the speakers' information. Hence, it was not always possible to know whether the analysed utterances were original, were produced by a native or non-native speaker, or came from a translated text. This might be potentially problematic when studying natural language use, in particular with variable article use. As mentioned before, the main purpose of the present case study is to examine whether translations can be used to investigate variable article use in English. This will be achieved by comparing a set of parallel texts that consists of English originals and German translations, with a dataset where the direction is not considered; put differently, the source language is not controlled for in the retrieval process.

Translations are often defined as a process that connects two or more languages and they can be very useful to analyse differences and/or similarities among language pairs. However, when investigating linguistic phenomena, there are a few questions that need to be addressed. The first point is whether we can trust translated texts to study a language. Two further points are whether translations are representative of the target language and whether they represent real language. These issues have been debated by many scholars. Some believe that translations cannot express the typical peculiarities of the language they portray. In translation studies, as indicated by Volansky et al. (2015: 3), research has argued that translated texts differ from original, non-translated material. Teich (2003: 20), for instance, affirms that "translations are a kind of text in its own right that has specific properties distinct from texts that are not translations". If this were the case, using translations as evidence for linguistic analysis would be considerably risky. Teubert's (1996: 247) firm opinion on translated texts goes in the same direction:

> Translations, however good and near-perfect they may be (but rarely are), cannot but give a distorted picture of the language they represent. Linguists should never rely on translations when they are describing a language. […] Rather than representing the language they are written in, they give a mirror image of their source language.

The relative importance of translations has been subject to considerable discussion. There are, in fact, other researchers who share the opposite viewpoint. Baker (1993: 234) states that the "traditional view of translation implies, in itself, an acknowledgment of the fact that translational behaviour is different from other types of linguistic behaviour, quite irrespective of the translator's mastery of the target

language." Put differently, according to Baker (1993), translated texts are not better or worse than original texts; they are simply different. This observation is shared by Mauranen (1999: 181), who claims that translations are texts functioning in the target language, "they assume real functions as texts in a living culture, not necessarily identical to those of the source texts in their original or other cultural contexts." In short, both Baker (1993) and Mauranen (1999) believe that translated texts have no negative connotation; on the contrary, they have the same autonomy and freedom of an original text.

When working with *Europarl*, however, it is important to mention two further obstacles that have to be faced. Firstly, it is possible that a small proportion of the texts contained in the corpus were translated by non-native speakers. According to Pym et al. (2013: 12), translators need to be highly skilled and qualified, and the information provided by the European Parliament website states that, ideally, translations are done by native speakers. Nevertheless, "enlargements have generated a powerful push towards greater efficiency in the operation of multilingualism. For example, the general rule that translators and interpreters work only into their mother tongue is slowly transforming" (*European Parliament – never lost in translation*, OD[49]). Using translations that are not done by native speakers might in turn represent a limitation for the study of natural language use. Unfortunately, information on translators' background is not available. The second additional challenge concerns the distinction between the first language and second language speech production. It is well established that parliamentary members, whose first language is not English, often prefer communicating in English rather using their mother tongue. In other words, untranslated English texts may be based on the utterances of non-native speakers of English.[50] This is also confirmed by Codrea-Rado (*The Guardian*, 21 May 2014, OD[51]), who observes that from 2008 to 2012 English was the most frequently used language at the plenary debates.

The current analysis builds on a contrastive study between English and German article use with collective nouns (discussed in Chapter 5). The present case

[49] Available at: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+IM-PRESS+20071017FCS11816+0+DOC+XML+V0//EN

[50] Note that, however, non-native speakers who nevertheless choose to use English in Parliament are likely to be rather highly competent speakers of English.

[51] Available at: http://www.theguardian.com/education/datablog/2014/may/21/european-parliament-english-language-official-debates-data

study takes into consideration the same nouns, i.e. *Parliament*, *Council*, *Committee*, and *people*. In addition, the data sample comes from the same corpus, i.e. the transcripts of the parliamentary speeches held at the European Parliament. The only essential difference between these two studies is the fact that the data are collected differently. In the present study, all instances are randomly retrieved from the corpus, without paying attention to the text type. Thus, no distinction between original texts and translated texts is made for the present study (this counts for both the English and German datasets). Furthermore, the analysis does not consider the speaker's first language. The results of the current investigation are then compared with the results of the analysis that will be discussed in more detail in Chapter 5, hereafter called *Study A*. The main aim of this research is to determine whether a case study with only English original texts (produced by English native speakers) and a case study with a combination of original texts (coming from politicians who might also be English non-native speakers) and translations produce different results. Additionally, it will help to investigate whether, to study natural article use, it is essential to analyse data that exclusively come from original texts and not from translations. Finally, this study seeks to examine whether translations can represent natural article use without being a result of first language influence.

The following section will introduce the theoretical background on translation studies and will focus on *translationese*, translation universals, and translations in relation to Corpus Linguistics. Section 4.3.2 will explain what data have been included in the analysis, and how they have been retrieved and annotated. Section 4.3.3 will show and analyse the results of the four collective nouns and compare them with the findings of Study A.

4.3.1 Translationese and (Parallel) Corpus Linguistics

Several attempts have been made to define translation language. As already mentioned, researchers have conflicting viewpoints: some linguists argue that translated texts should not be included in a corpus (e.g. Teubert 1996); others (e.g. Baker 1993, Mauranen 1999) claim that they are as acceptable as original texts. A key aspect of this topic is how translated texts have slowly acquired importance and independence. In other words, there has been a shift from source language oriented studies to target language oriented studies. Initially, translations tended to be seen in a

negative light. The comparison between translated texts and original texts led to translations being considered inadequate and unfaithful, because of their inability to perfectly represent the source language's features. At a later stage, scholars began to gradually idealise translations, which then started to have a more positive connotation. Translated texts were no longer strictly related to the source text. Additionally, notions like *equivalence* and *correspondence* were brought to the fore (Baker 1993: 235-236, Puurtinen 2003: 391). Attention moved from the primacy of the source text to the qualities of the target text. For this reason, the term *traslationese*, a locution to determine translated texts' qualities, is now used without negative connotations and simply refers to the specific language of translations (Puurtinen 2003: 391).

However, Tirkkonen-Condit (2002: 207) makes a clear distinction between the translationese of *bad* translations, and translationese in general, whose "linguistic or textual features in translated texts […] cannot be avoided." Moreover, Puurtinen (2003: 391) notes that "[t]ranslationese may be the result of source language interference, which has led to the use of target language forms and structures which are formally equivalent to some source language forms and structures." Theorists have then indicated some particular characteristics of translated texts, which are commonly called *translation universals*. These are generally defined as typical and unavoidable tendencies caused by the translation itself. In other words, these tendencies "do not relate to translation errors but to frequencies of lexical items, syntactic patterns, etc. which deviate from those in originally produced texts" (Tirkkonen-Condit 2002: 208). The following are the most important universal features of translations highlighted by Baker (1993: 243-244):

> (i) A marked rise in the level of explicitness compared to specific source texts and to original texts in general. […] (ii) A tendency towards disambiguation and simplification. (iii) A strong preference for conventional 'grammaticality'. […] (iv) A tendency to avoid repetitions which occur in source texts, either by omitting them or rewording them. […] (v) A general tendency to exaggerate features of the target language.

Hence, the tendencies reflected in translations can be summarized as the following: *simplification*, *normalization*, *explicitness*, and *general conservatism* (see Baker 1993, 1995, 1996, Laviosa-Braithwaite 1996, Tirkkonen-Condit 2002). It is worth noting that, with respect to normalisation, Tirkkonen-Condit (2002: 208) suggests the so-called *normalcy hypothesis*, which "predicts that translations tend to exaggerate those

features that are frequent in the target language, and that metaphors and idioms would be more conventional, and dialectal and colloquial expressions less frequent."

It is clear that, in linguistic analysis, using texts produced by non-native speakers entails certain limitations. To illustrate, when passing from L1 to L2, there is a high risk of the so-called *cross-linguistic influence*, alternatively called *language transfer* (Odlin 1989: 1), which includes phenomena such as *transfer*, *interference*, *avoidance*, and *borrowing* (Sharwood Smith and Kellerman 1986: 1). In other words, elements of a source language (i.e. L1) are sometimes (positively or negatively) transferred into a target language (i.e. L2 or L3)[52], which lead to an oral or written production/performance which might not successfully represent natural language. There is therefore the chance that these features can interfere in the results of the linguistic investigation.

In the last decade, translationese has been attracting a lot of interest in different linguistic research fields, especially in (Parallel) Corpus Linguistics (see e.g. Baroni and Bernardini 2006, van Halteren 2008, Kurokawa et al. 2009, Ilisei et al. 2010, Koppel and Ordan 2011, Lembersky et al. 2012, 2013, Twitto-Shmuel et al. 2015, Volansky et al. 2015). In particular, in more recent studies, Rabinovich et al. (2016) and Nisioi et al. (2016) used *Europarl* to investigate the main features of translations in comparison to original texts. In order to do so, they obtained three sub-corpora, i.e. one with instances uttered by native English speakers, one by non-native English speakers, and one with English translations from various European languages, and systematically compared them. The most relevant findings of these studies indicate that "native texts are easily distinguishable from the other two classes" (Nisioi et al. 2016: 4199), and that "[t]here are clear similarities between translations and non-native language" (Rabinovich et al. 2016: 1870). In particular, Nisioi et al. (2016)'s results exhibit that the type-to-token ratio (TTR) is much higher in native productions than in translated texts and non-native texts. Put differently, "translated texts tend to exhibit less lexical diversity and vocabulary richness" (Nisioi et al. 2016: 4199). Interestingly, the fact that TTR of non-native productions is lower than the one of translated texts reflects that "the lexical diversity of (highly competent) non-native speakers is poorer than that of translations, who translate into their mother tongue" (Nisioi et al. 2016: 4200). These results were later confirmed in the study conducted

[52] Note that a transfer can also happen from L2 to L3 (Sharwood Smith and Kellerman 1986: 1, Odlin 1989: 27).

by Rabinovich et al. (2016). In native productions, they find a higher use of personal pronouns and collocations (e.g. *bring forward, food chain*) and a lower use of transition markers (e.g. *in addition, at the same time, thus*).

Based on the results of the above-mentioned studies, it seems that, overall, non-native and translated texts are more similar to each other and in contrast with native speakers' productions, which tend to be easier to identify. In another study, Bernardini et al. (2016) do not exclusively take into account translations but mainly focus on the comparison between translation and interpreting. Their analysis is an attempt at building an *intermodal corpus*, using EPIC (the European Parliament Interpreting Corpus) as their starting point. Intermodal corpora are generally defined as "corpora containing parallel or comparable outputs of translation and interpreting" (Bernardini et al. 2016: 2). The resulting corpus is called EPTIC (the European Parliament Translation and Interpreting Corpus), a bilingual and bidirectional corpus of English and Italian, with which lexical simplification is investigated.[53] The main findings show that translated texts are more complex than interpreted texts. On a monolingual comparable level, the mediated texts are simpler than the non-mediated ones. It seems, however, that "different parameters of simplification apply differently to the two languages" (Bernardini et al. 2016: 20). Taken together, the findings indicate that the input used by interpreters is simpler than the one used by translators; put differently, "spoken language is simpler than written language" (Bernardini et al. 2016: 19).

The studies presented thus far provide evidence that translations differ from original texts and indirectly support the argumentations discussed by e.g. Teubert (1996), Teich (2003) and Volansky et al. (2015), who claim that translated material should not be considered when investigating a language, because translations cannot fully represent their target language. As Teich (2003: 219) clearly states, "what makes translations different from original texts in the same language as the target language is that the source language shines through in translations."

[53] The corpus contains nine sub-corpora of both source and target texts, namely three of source speeches and six of interpreted speeches (Bernardini et al. 2016: 8).

4.3.2 Preliminary case study: data retrieval and annotation

The data sample used for this investigation comes from the first version of *CoStEp* (Graën, Batinic and Volk 2014). In Study A, the main focus is to exclusively work with the original texts, in order to analyse natural article use in English. Therefore, for the English dataset it is essential to retrieve only those instances that were originally produced by English native speakers. Thus, the data sample contains aligned sentences with the original text in English and the corresponding translation in German. By contrast, the present comparative study makes use of a different methodology for the investigation of article use with collectives in English and German. This time, no distinction is made between original texts and translations. In other words, the English and German datasets are a mixture of both original and translated texts. The results of the current analysis will be compared with the results of Study A. The comparison will help to understand whether the investigation of English article use needs to be restricted to text produced by English native speakers, or whether it is possible to make use of linguistic analyses whose data contain both original texts and translations.

For the data retrieval, no attention was paid to the speakers and their first language. The German dataset includes the parallel sentences of the English sample, and might also contain original texts and translations. The final number of parallel sentences of the data sample used for the investigation is 1.668 (853 in English, 814 in German).[54] The difference in the number of the English and German instances is due to wrong alignments that can occur. In these cases, the erroneously aligned German translations were excluded from the analysis.

When comparing the results of different case studies, it is important to use the same annotation system. The English and the German datasets are annotated manually. The annotation of Study A concerns the article type, the modification of the noun, its syntactic function, and the subject-verb agreement. These factors are chosen because they allow to investigate the extent to which article use might be influenced. By contrast, since the main goal of the present investigation is the comparison of article distribution using two different data retrieval methods, the analysis considers

[54] Note that, for the present study, all cases of *Parliament*, *Council*, and *Committee* were retrieved with upper-case, and that all cases of *people*, on the contrary, were retrieved with lower-case.

only the annotation of the article and does not include the annotation of the noun pre- and postmodification, the syntactic function, and subject-verb agreement.

There are three article types: definite, indefinite, and no article. In some cases, instead of an article, a different determiner is used, for example a demonstrative or a possessive; these cases are annotated as 'other'. As mentioned above, the results presented in the current study do not show the distribution of the noun modification. However, in order to have comparable data, the sample has to include collectives with the same type of modification of the data of Study A. The premodification considers either a singular or a plural common noun (e.g. *the Agriculture Council*), an adjective (e.g. *das Europäische Parlament*), a proper noun (e.g. *the Euro Council*), an acronym (e.g. *the Ecofin Council*), a number (e.g. *the 133 Committee*), or a combination of them (e.g. *the Internal Market Council, the Legal Affairs Committee*). Contrary to Study A, due to the low frequency, German compound nouns (e.g. *das Überwachungskomitee*) are considered as premodifiers. On the other hand, prepositional phrases are the only postmodifiers (e.g. *the Committee on Foreign Affairs, Human Rights, Common Security and Defence Policy*). Similarly to Study A, the German dataset also includes phrases marked with a genitive case (e.g. *das Internationale Komitee des Roten Kreuzes*). Finally, it is worth noting that the present data comprise many cases with both premodification and postmodification (e.g. *the International Committee of the Red Cross*, *the Helsinki Committee for Human Rights* or *the European Committee for Standardisation*). This has to be taken into account for the analysis, as this constitutes a marked difference to Study A, which does not include similar cases. For this reason, cases of this type were deleted from the dataset.

The present study takes into consideration the cases where the collective is the head noun and not an element which modifies a different head noun. Some examples for constructions that are not part of the analysis – i.e. a first type of false positive – are found in (1) – (4).

(1) […] means in fact that the 1982 Council regulation on the implementation of the Convention is in urgent […]. (*CoStEp*, 1996-09-17.xml)
(2) I am not speaking in my capacity as Committee Chairman but as rapporteur for […]. (*CoStEp*, 2007-11-12.xml)
(3) […] to strengthen the institutional framework of the International Accounting Standards Committee (IASC) Foundation. (*CoStEp*, 2007-11-12.xml)
(4) Let us rely on the Council (or all the Council members) to look at it in this way. (*CoStEp*, 1996-11-12.xml)

Two further types of false positives are the cases where the collective noun is used as an event, as in (5), and genitive cases, as in (6), (7) and (8), because the article does not refer to the collective noun.

(5) […] a country with whom it was possible to announce, during the General Affairs Council last Monday, that the negotiations led by […]. (*CoStEp*, 1997-09-17.xml)
(6) […] given that the Council's obligation of transparency differs according to […]. (*CoStEp*, 1996-07-17.xml)
(7) […] referring to the present Chairman of the Russian Social Democratic People's Party […]. (*CoStEp*, 2002-09-26.xml)
(8) […] with textile uppers originating in the People's Republic of China and Indonesia. (*CoStEp*, 1996-12-10.xml)

In the following, the description of the results and the comparisons between translated and original material are given.

4.3.3 Preliminary case study: results and analysis

A. PARLIAMENT

The first collective noun is *Parliament* in English and *Parlament* in German. It is worth noting that the English part of the data included false positives like in (9) and (10).

(9) My group is amazed that the signatories to this motion, so particular about the rights of the Members of the European Parliament […]. (*CoStEp*, 1997-02-20.xml)
(10) I believe we in the European Parliament must show our utter condemnation of these atrocities […]. (*CoStEp*, 1996-10-24.xml)

These cases and their equivalents in German had to be excluded from the dataset, because these are not instances of the collective noun, but they are part of a more complex noun phrase where the noun *Parliament* is used to refer to the institution. In addition, German sometimes uses an acronym instead of a phrasal expression, i.e. *das Europäische Parlament* is substituted by *das EP*. This choice is probably due to the necessity to type fast, and, therefore, might constitute a typical feature of a translation as opposed to an original text.

Figure 4.8 and Figure 4.9 compare the results of the current study with Study A in English and German, respectively. In the samples where the speaker nationality and the source language are not controlled for, it is possible to see that English and German have many aspects in common, and that the overall results are in fact very similar. Thus, in both languages the definite article is used more frequently than the bare NPs and other determiners. Additionally, the frequency of the indefinite article is extremely low. However, English slightly differs from German in respect of variability. In fact, while German clearly shows a preference for the definite article, English varies between the definite and no article.

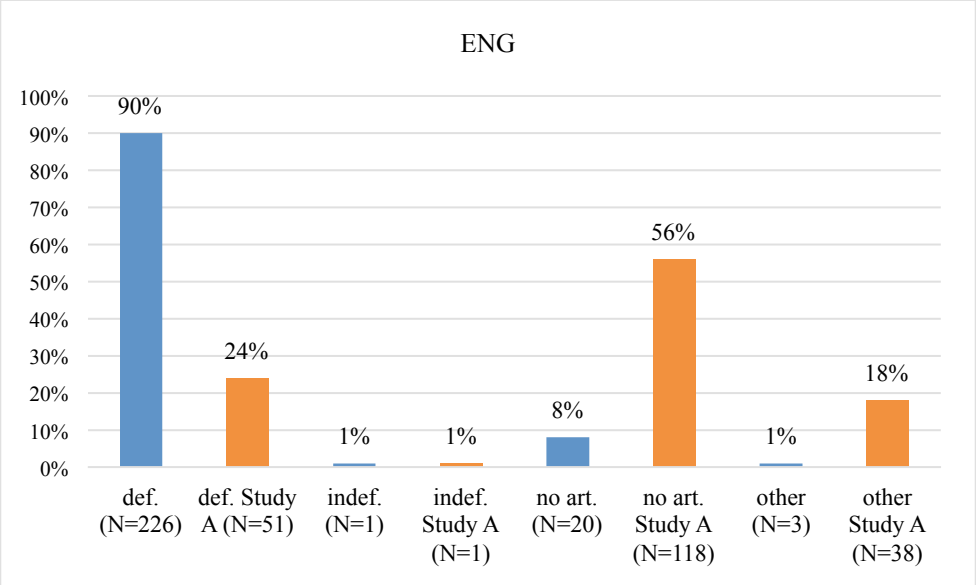

Figure 4.8: Comparison between current study and Study A for *Parliament* in English.

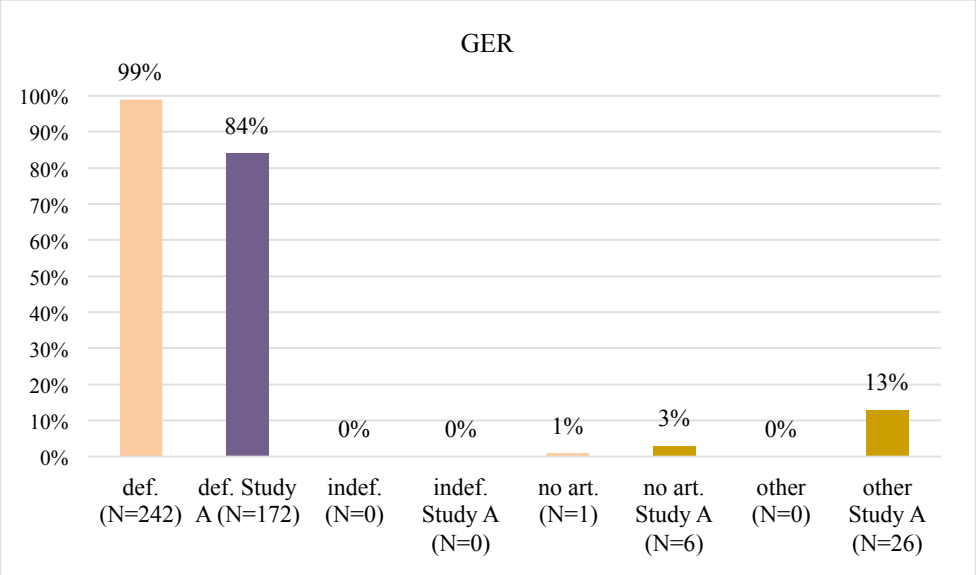

Figure 4.9: Comparison between current study and Study A for *Parliament* in German.

Turning now to the comparison between the present analysis and Study A, one can note that the results do not perfectly match. Figure 4.8 puts in contrast the findings in English. What is striking in this graph is that the sample that controls for source language and native-speakerhood shows more variability between article use and article omission. In other words, Study A produces an overall higher number of bare NP uses. However, a closer look at the present data reveals that the high frequency of cases occurring with the definite article is due to the premodifying elements (i.e. the expression *the European Parliament* occurs more often than simply *Parliament*). As will be discussed in more detail in Chapter 5, premodifiers can play a relevant role in article use; they in fact make the whole NP more specific, which in turn favours the use of the definite article. The preference of using a modifier in front of *Parliament* in translations might be interpreted as an indication of non-original texts. A further characteristic that allows us to speculate that the sentence shown in (11) was uttered by non-native English speaker is the use of the definite article with *Parliament*. This is the only case in which this collective occurs with an article and has no modification.[55]

[55] At a later stage of the project, it was possible to confirm that the reported instance is a translation; the original sentence was uttered by a Portuguese native speaker.

(11) The Parliament should therefore give a [sic] it a positive signal that will make it easy for the public to […]. (*CoStEp*, 1997-11-19.xml)

The results for German are shown in Figure 4.9. Even though the German translations of the English sentences also have a slightly higher use of bare NPs, the findings exhibit stronger similarities between the two studies, namely a solid preference towards the use of the definite article.

B. COUNCIL

The second noun is the English collective *Council*. The English subset of the present study included some instances that could not be taken into account in the analysis. Two examples are shown in (12) and (13).

(12) Mr. President-in-Office of the Council, you made a reference to coastal patrol […]. (*CoStEp*, 2008-05-21.xml)
(13) […] in this context the EU needs reliable consumers' organisations, indeed organisations like the Danish Consumers' Advisory Council, to […]. (*CoStEp*, 1998-10-07.xml)

These cases were excluded because the noun *Council* refers to an institution rather than the collective of the individuals.

Figure 4.10 and Figure 4.11 combine the results of the current study with the findings of Study A in English and German, respectively. When comparing the results of the present analysis between English and German, it is possible to note that both languages never use the indefinite article, and that there is only one case in English where a different determiner is used instead of an article.

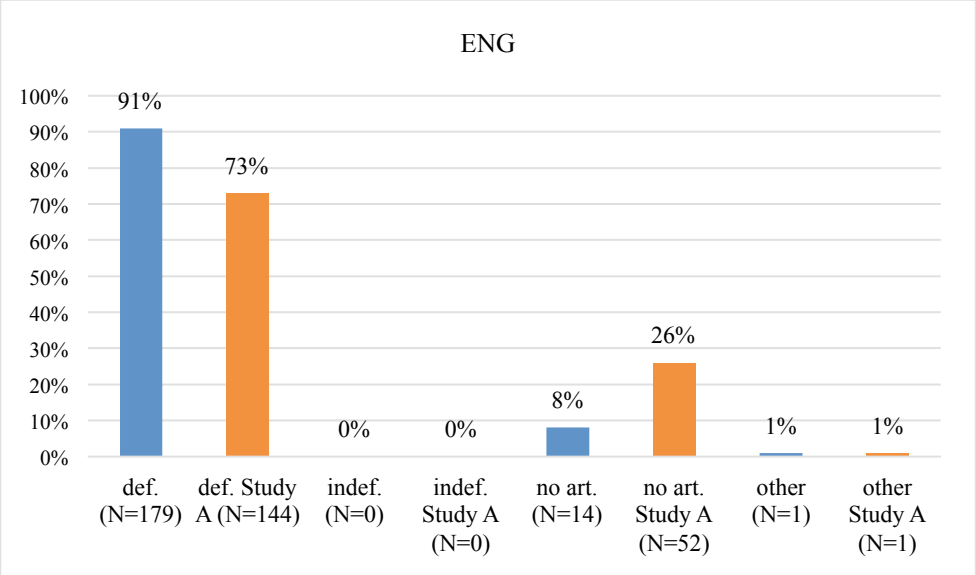

Figure 4.10: Comparison between current study and Study A for *Council* in English.

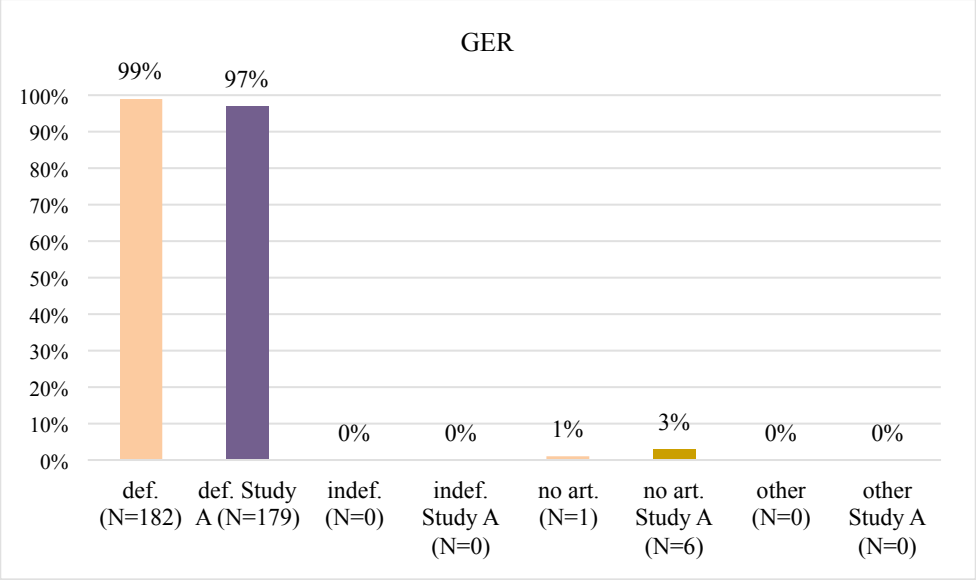

Figure 4.11: Comparison between current study and Study A for *Council* in German.

The results make it clear that both English and German highly prefer the definite article (more than 90% of the data). This preference is categorical in German, as the article is omitted in only one utterance of the sample. By contrast, English article use is more variable, albeit at a low rate; articles are, in fact, omitted in 8% of the data.

Let us now consider the comparison between the current study and Study A. Figure 4.10 shows that in English the results of the two analyses are not comparable.

Similarly to the previous collective noun, the current study exhibits a higher number of cases in which the definite article is used. Including texts that are translations or those that were not produced by native speakers skews the sample towards fewer bare NPs. On the other hand, as shown in Figure 4.11, German shares more similarities between the two studies, i.e. the definite article occurs almost categorically, while the frequencies of the indefinite article, bare cases and other determiners are extremely low. Once again, Study A presents a few more cases of bare NPs, these will be analysed in more depth in Chapter 5.

C. COMMITTEE

The noun *Committee* is the third collective considered in the present case study. Contrary to Study A, there is not only one single German version, i.e. *Ausschuss*; rather, the noun *Komitee* is also part of the data. Figure 4.12 and Figure 4.13 present the results of English and German in the present investigation and compare them with Study A. In the current analysis, the findings show that in both English and German the indefinite article is hardly ever used. In addition, in some utterances German uses other determiners more often than English. The results also show that both languages favour the use of the definite article. However, this preference is stronger in English than in German. A further difference that needs to be pointed out regards the distributions of article omission. In English, there are no bare cases, while German shows a relatively high frequency. A closer look at the data shows that more than half of these cases occur with postmodifiers. Furthermore, these instances are used as bare NPs because the noun is either part of a coordinated subject[56], as in (14), or follows *als*, as in (15), which does not require an article in German.

(14) [...] wie dem Assoziationsrat und Assoziationsausschuss […]. (*CoStEp*, 2002-03-13.xml)
(15) Wir werden dann unsere Verantwortung als Ausschuss für Beschäftigung und soziale Angelegenheiten übernehmen […]. (*CoStEp*, 2009-12-14.xml)

[56] Coordinated phrases can be difficult cases when investigating article variability, because it is not always possible to know whether the article preceding the first element in a coordinated structure is within the scope of the whole NP. However, since an article could be used but is not pronounced by the speaker, these cases have been included in the datasets as well.

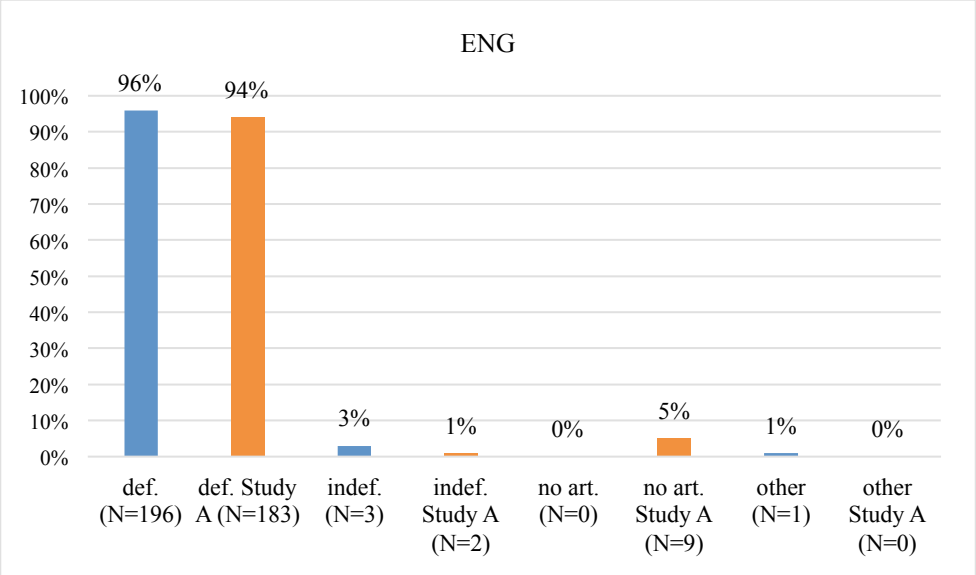

Figure 4.12: Comparison between current study and Study A for *Committee* in English.

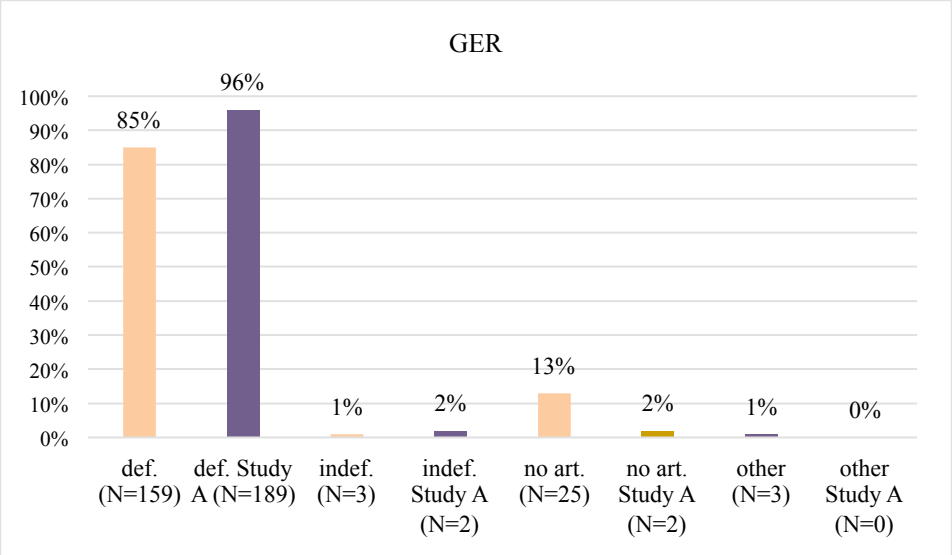

Figure 4.13: Comparison between current study and Study A for *Committee* in German.

Turning now to the comparison between the two retrieval methodologies, it is possible to see that English, shown in Figure 4.12, exhibits similar tendencies. In both investigations, the definite article is highly preferred. As will be shown in the analysis of Study A, the strong tendency towards article use is influenced by the modification of the noun. However, once again, the controlled-for dataset consistently yields higher numbers of bare NPs. On the contrary, Figure 4.13 shows that in German there is a substantial difference between the two studies, i.e. in the present data, German

displays more cases with article omission. However, as pointed out before, these are not to be considered as genuine bare NPs.

D. PEOPLE

The final collective is *people*. As will be shown in Study A, this noun in German has more lexical variability. In other words, it is translated in many different ways (e.g. *Behinderte(n)*, *Bevölkerung*, *Verbraucher*, or *Volk*), and therefore, there is more specificity in regards to different categories of *people*. Figure 4.14 and Figure 4.15 provide the results obtained from the analysis in English and German, together with the findings of Study A. With respect to the comparison between the two languages in the current study, the results show that, on the one hand, they share some limited similarities but, on the other hand, generally differ widely from each other. Firstly, there are no cases in which the indefinite article or another determiner is used. This, however, is not surprising, as the noun *people* is chiefly used in the plural form, which would therefore conflict with the use of the indefinite article. These characteristics perfectly represent the equivalence between the two languages. However, as previously mentioned, the results exhibit a considerable difference, i.e. bare NPs are very frequent in English, whereas in German the definite article is highly favoured.

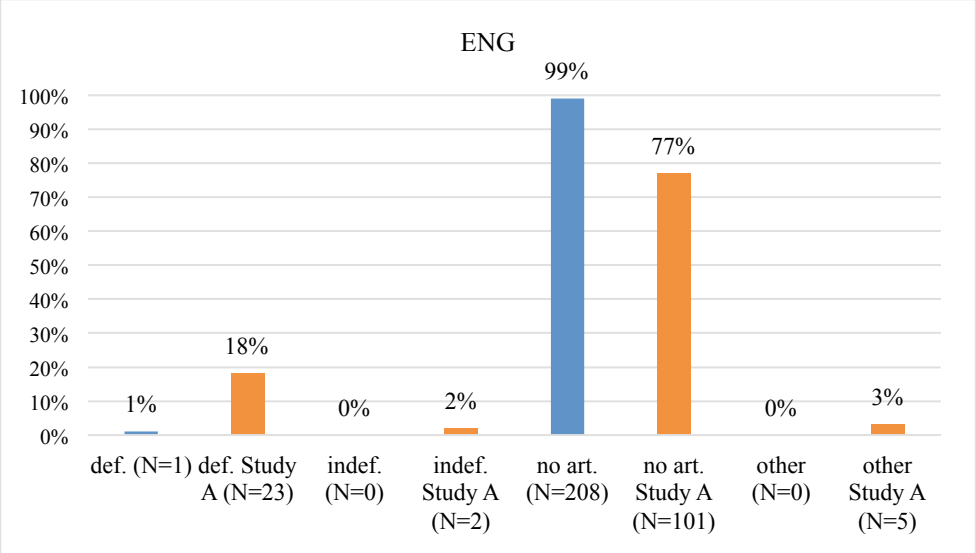

Figure 4.14: Comparison between current study and Study A for *people* in English.

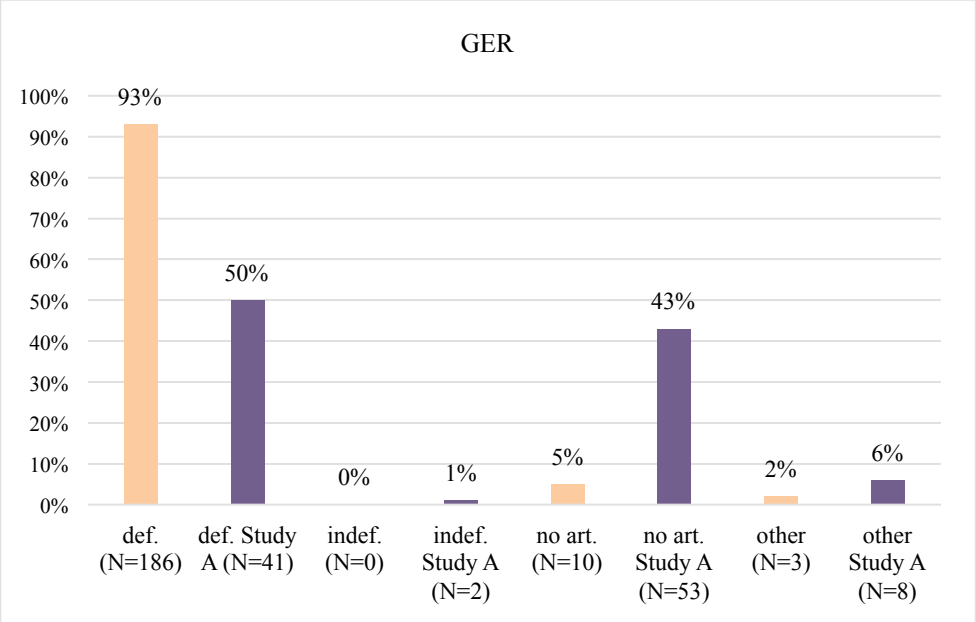

Figure 4.15: Comparison between current study and Study A for *people* in German.

The comparison between the current study and Study A indicates that the results are partly different. In both languages, Study A presents more cases in which another determiner is used instead of an article. Furthermore, the most striking findings in English, shown in Figure 4.14, is the fact that in the sample including translations and non-native English speakers there is a lower use of the definite article and articles are omitted more often than Study A. On the other hand, Figure 4.15 shows that, in Study A, German shows a marked variability between the article use and article omission, which is not attested in the present study. This marked dissimilarity may be due to the type of parallel sentences that were randomly retrieved for the investigation. In fact, by looking at the data more closely, one can note that the collective noun appears in contexts that are very similar to each other. Cases like the ones shown in (16) – (19) often occur.

(16) *People* no longer have to prove that they are superior to machines. (*CoStEp*, 1997-09-15.xml)
(17) *Der Mensch* muß nicht mehr beweisen, daß seine Leistungen einer Maschine überlegen wären. (*CoStEp*, 1997-09-15.xml)
(18) *People* are rising up against dictatorial despots and that is a good thing. (*CoStEp*, 2011-02-02.xml)
(19) *Das Volk* erhebt sich gegen diktatorische Despoten, und das ist gut so. (*CoStEp*, 2011-02-02.xml)

In both English examples, *people* is the subject of the instance and has no article, whereas the German equivalents – i.e *Der Mensch* and *Die Bewohner* – are also in subject position but are used with the definite article.

The important key points of the two discussed preliminary case studies are summarized in the following section.

4.4 Summary

One of the more significant findings emerging from the first preliminary study (discussed in section 4.2) is that there are some discrepancies between the debates' transcripts and their equivalent video recordings in both English and German. Furthermore, the analysis has shown that English has more cases of dissimilarities concerning the addition or the omission of an article. Hence, the results confirm the initial expectations, namely that articles in English are more likely to be omitted than in German, which, by contrast, has a stronger tendency to be more conservative with regard to grammar rules. However, the analysis of the changes adopted in the transcripts revealed that in both English and German the majority of cases are due to variability and not to grammaticality. This means that the transcribers in both languages tend to use more traditional forms, because the constructions used by the speakers are not well established in the written medium. Additionally, this study has also identified that English and German share other similarities. One of the major findings was that in both groups the differences concern more singular cases and NPs that refer to abstract nouns. In English, this category is followed by institutional nouns.

Articles		
	ENG	**GER**
	Indefinite article	Indefinite article
11 Dec 12	281	302
1 Jul 13	99	142
21 Oct 14	108	162
Total	*488*	*606*
	Definite article	Definite article
11 Dec 12	778	1.183
1 Jul 13	319	480
21 Oct 14	354	899
Total	*1.451*	*2.562*
	-	Article within a preposition
11 Dec 12	-	185
1 Jul 13	-	91
21 Oct 14	-	107
Total	-	*383*
Sum	***1.939***	***3.551***

Table 4.8: Comparison between English and German of articles occurring in the three transcripts (raw numbers).

Finally, to better define the *Europarl* corpus' reliability, it is important to have a general overview of article use. Table 4.8 shows the raw number of all articles occurring in the three transcripts in both English and German, while Table 4.9 gives the total number of bare NPs.

Bare NPs		
	ENG	**GER**
11 Dec 12	1.397	1.465
1 Jul 13	698	751
21 Oct 14	536	1.021
Total	***2.631***	***3.237***

Table 4.9: Comparison between English and German of bare NPs occurring in the three transcripts (raw numbers).

The number of the analyzed turns amounts to 260, and the number of differences between the speakers' speeches and the verbatim of the proceedings are 75 for English and 28 for German. Thus, the discrepancies between the politicians' original speeches and the parliamentary transcripts correspond to 4% for English and only 1%

for German.[57] In general, therefore, the transcripts of the European Parliament are very faithful to what is said by the parliamentary members during the sittings. Since the study was limited to article use, it was not possible to examine the transcripts in their entirety. Notwithstanding this limitation, the study suggests that the corpus used for the current project is reliable for the investigation of article use. Moreover, it provides additional evidence with respect to the relation between written and spoken language, with particular reference to the written language of spoken discourse in an institutional context.

The second preliminary case study (discussed in section 4.3) tested whether data comprised of both original texts (i.e. source-language English produced by native speakers) and translations provide a reliable basis for the investigation of natural article use in English. It reproduced a contrastive study of article use with collective nouns (i.e. Study A, discussed in detail in Chapter 5), and finally compared the results. The essential difference between the two contrastive studies concerned the data retrieval method. Study A based its examination on utterances exclusively produced by English native speakers, whereas the present study used mixed instances of both originals and translations. The analysed collectives were *Parliament*, *Council*, *Committee*, and *people*. Taken together, the results of this study have shown that including in the analysis translated material and texts that were not produced by English native speakers yields a lower number of bare NPs. Particularly in English, the sample where the source language and the speakers' mother tongue were controlled for produced more bare NP uses. The only collective noun in which this was not confirmed was *people*. Furthermore, the German findings of the current analysis have also shown a lower use of bare NPs. However, many of these cases were not genuine bare NPs. Overall, the investigation has shown that the results of the current study differ from the ones of Study A with respect to the frequency of bare NPs. Hence, the results of this investigation complement those of earlier studies (e.g. Nisioi et al. 2016, Rabinovich et al. 2016, and Bernardini et al. 2016) and strengthen the notion that translations deviate from original material. Furthermore, they support the idea that a data sample including both original texts (produced by both English native and non-native speakers) and translations is thus not fully reliable for the investigation of natural article use in English. Hence, the current data highlight the

[57] Based on the number of articles in the transcripts shown in Table 4.7.

importance of the use of original texts for linguistic investigation and make clear that the current project cannot take into consideration mixed texts for the target language.

A final observation that also needs to be pointed out regards the methodology used in the research project. As already mentioned, the aim is to model natural language use in English by using German translations as a starting point. The second case study has already shown that the German sample with translations from different languages slightly differs from the sample of Study A, which only includes translations from English original texts. This shows that the language used as starting point can somehow influence the final results. Put differently, the multi-lingual perspective used to target original bare NPs in English might therefore show a bias in terms of the data that will be retrieved. This is thus a limitation that will have to be taken into consideration; however, this can also be a stimulating point for further research. With the following chapter, the second part of this dissertation begins. Chapter 5 will look in more detail at Study A, which uses English original material to investigate article use with collective nouns and compares the results with German. The following contrastive study will be valuable for the evaluation of the opportunities that this parallel corpus provides.

Part II – Case studies

5 The lexeme based-approach: collective nouns

5.1 Introduction

Linguistic comparison is a central component of linguistics (Willems et al. 2004: 1). Many researchers have worked to study the use of articles in both English and German, but not many studies have been conducted to analyse these two languages in contrast. The following case study is a contrastive analysis and compares English and German. It is the first analysis on article use that was conducted in this research project, and the data used for this investigation come from the first cleaned version of the *Europarl* corpus, i.e. *CoStEP*. As already seen in Chapter 2, the existing body of research suggests that both English and German standard grammars agree on the concepts of the various articles; namely that the definite article is used in front of noun phrases to express definiteness, while the indefinite article expresses indefiniteness. However, article omission gives a conception of a whole class with a general connotation (Curme 1970: 62, Quirk et al. 1985: 282, Biber et al. 1999: 261, Huddleston and Pullum 2002: 407, Dudenredaktion 2005: 338). The current investigation is thus a study which is useful to prove that corpus evidence can provide new insights into both (Parallel) Corpus Linguistics and Contrastive Linguistics. Moreover, it provides an important opportunity to advance our understanding of the differences and similarities between English and German in relation to article use.

As already mentioned in section 4.3, the current contrastive study is lexeme-based and its main aim is to investigate the differences in article use with collectives in English and German. At the beginning of the current project, a preliminary randomized data sample containing English and German parallel sentences was retrieved for alignment evaluation. Based on a qualitative investigation of the data sample, article variability was noticed with the noun *Parliament*. Two frequency lists with *Parliament* occurring with and without an article were retrieved. The comparison of the lists confirmed that *Parliament* showed variability with respect to article use (i.e. the expression *the Parliament* occurred 934 times, while the bare NP *Parliament* occurred 25.791 times). Since *Europarl* mainly contains parliamentary language,

collective nouns are expected to commonly occur and are therefore worth investigating in more detail. Based on a frequency list containing nouns > 10 in the English part of the corpus, a small set of collective nouns with a relatively high frequency were selected. The nouns taken into consideration are *Parliament*, *Council*, *Committee*, and *people*. The data used in this case study consist of original English sentences and their parallel German equivalents. In other words, sentences uttered by native English speakers and their corresponding German translations. The first question that this analysis will try to answer is whether articles preceding collective nouns are used in the same way in English and German and what the differences are. In addition, article use in both languages can be influenced by various factors; one of the possible factors might be the noun's syntactic function. Therefore, the study will try to determine any article variability with collectives in subject and non-subject position. Moreover, another interesting context is the subject and verb agreement (alternatively called *concord*); as Depraetere (2003: 86) points out, "[…] the system of concord is variable" and it is, therefore, worth investigating whether the concord might influence article use, too. Finally, as already described in the theoretical part of this dissertation, BrE and IrE greatly differ concerning article use, since IrE uses the definite article more frequently than BrE (see for instance Hickey 2007: 251, Corrigan 2010: 52, and Kallen 2013: 122). Based on this statement, the present study will investigate whether there are any differences (or similarities) between British and Irish speakers as regards article use with collectives.

The analysis is organized in the following way. Section 5.2 will give a general overview on the use of articles with collective nouns and on the relationship between concord and collectives. Section 5.3 will provide information regarding the methodology used for the investigation, the data retrieval, and data annotation. Section 5.4 will present the results of the analysis of the articles with every single collective, the relation with the syntactic function and with the subject-verb agreement, the article distribution between British and Irish speakers, and the logistic regression analysis. Finally, section 5.5 will summarize the results and will conclude the chapter.

5.2 Theoretical background

5.2.1 Article use with collective nouns

Crystal (2008: 86) defines a collective as "a noun which denotes a group of entities, and which is formally differentiated from other nouns by a distinct pattern of number contrast." This definition is largely shared by many other scholars and is discussed, for instance, in Depraetere (2003: 86) and Levin (2006: 321). Unfortunately, research has not provided much information about the use of articles with collectives, especially because it is quite unpredictable. This is confirmed by Christophersen (1939: 27), who states that: "so far as the articles go, there are no formal criteria for the recognition of a group of collectives." The definition given by the dictionaries implies that most of the collectives consist of smaller elements. But should *house* be considered a collective since it exists thanks to the combination of other elements, like walls, doors, and windows (Christophersen 1939: 161)? According to other researchers' arguments, it should. Depraetere (2003: 86), for instance, affirms that nouns like "train (a unity of wagons), forest (a set of trees), and luggage (a collection of suitcases) are to be classified as collective nouns as well." Previously, Jespersen (1949: 93) expressed the same concept, stating that these nouns are considered collectives due to the fact that they are singular but denote a collection or number of individuals. However, the difficulty lies in the fact that the definition sometimes is not straightforward.

Within the category of collectives, nouns can have many different properties. For instance, Kruisinga (1960: 62-63) distinguishes between collective nouns denoting people or the idea of the unity as a group (*e.g. Government, class, Cabinet, population*), personal collectives (*e.g. folk, cavalry, police*), and non-personal collectives (*e.g. information, hair, fruit*). In addition, Cruse (1986: 176) focuses on the differentiation between the group-member relation (*e.g. tribe, committee, family, orchestra, audience*), the class-member relation (*e.g. proletariat, clergy, aristocracy*), and the collection-member relation (*e.g. heap, forest, library*). Moreover, Christophersen (1939: 161) argues that there can be a distinction between the ones considered non-countable (*e.g. humanity, youth*) and the ones considered countable (*e.g. army, group, people*). It therefore seems that the use of articles with collectives is not to be considered regular, and he finally suggests (1939: 160) that: "[t]he only

criterion that we have to go upon is the meaning." Thus, according to Christophersen, the use of articles with collectives depends on what type of noun the collectives are part of. Poutsma (1904: 590-594) only analyses the group of collective nouns when referring to the definite article with generic reference, and also states that the use of the definite article is less regular. According to Poutsma, the definite article is used with collectives that express a class, a sect or a section of society (*e.g. aristocracy, community, democracy, mass, people, public*). By contrast, the definite article is omitted in front of nouns that have a comprehensive reference (*e.g. Christendom, humanity, manhood, royalty, society*) and in front of nouns that denote abstract, religious or artistic aspects (*e.g. childhood, youth, Catholicism, Paganism*). Hence, the original abstract meaning of these nouns explains the omission of the article.

Unfortunately, there is not enough information for the German language. Dudenredaktion (2005: 176-177), referring to collectives, does not mention article use but only observes that German countable collective nouns (*e.g. Gruppe, Herde*) can have the plural form, and that words like *Obst* (*i.e. fruit*) and *Gemüse* (*i.e. vegetables*) are not considered collectives, but simply non-countable nouns.

5.2.2 Collective nouns and subject-verb agreement

When investigating subject-verb agreement, the literature has mainly focused on the relationship between concord and the semantic aspect of collective nouns, without paying much attention to article use. According to Poutsma (1904: 277), concord refers to "[t]he way in which certain elements of a sentence, or a complex or clauses, are related, causes a certain analogy or agreement in number, person, gender and case […]." Collins Online (2017) indicates that

> [c]ollective nouns are usually used with singular verbs: *the family is on holiday*; *General Motors is mounting a big sales campaign*. In British usage, however, plural verbs are sometimes employed in this context, esp (sic) when reference is being made to a collection of individual objects or people rather than to the group as a unit: *the family are all on holiday*.

Bailey (1987: 3-6) tries to prove that in BrE there is a current change among younger speakers, who prefer to avoid plurals and to favour, instead, the singular form with collective nouns. On the contrary, Fries (1988: 103) is of the opinion that "young people [do not] in any way avoid plural forms, but much rather that they do not care

at all and use singulars and plurals indiscriminately […].” However, Levin (2006: 324) notes that in the nineteenth century BrE had a shift from the use of plural verbs to singular verbs with collective nouns. He continues: “[i]t therefore seems that most nouns in BrE have drifted towards singular verb agreement and that there are some which are resisting this trend”. Previously, Depraetere (2003) came to the same conclusion, attesting a preference for singular verbs. In the same study, she also focuses on the factors that might determine the agreement, and the determiners are one of them. However, she simply summarizes other scholars’ observations, namely, that collectives preceded by the indefinite article are obviously used with a singular verb, and that the definite article precedes those collectives exclusively used in the singular.

With respect to German, there seems to be a lack of research in this area. No studies have been found dealing with the influence of concord with collective nouns on article use.

5.3 Data and methodology

5.3.1 Data retrieval

Using parallel texts means that the analysed instances might come from a text which is either the original or a translation. It is, therefore, very important to determine the language of the original text and the language of its equivalent translation; furthermore, ensuring that the data sample is not composed of a mixture of both original and translated instances is required, too (see section 4.3). For the present analysis, the aligned sentences have their original version in English and their equivalent translation in German. A further limitation of this parallel corpus is that in the European Parliament there are many politicians who talk in a language that is not their mother tongue. This is particularly the case with English, because speakers are more likely to use English as a second language than German. Hence, being able to retrieve instances uttered by English native speakers together with their parallel translations in German was an important step for this analysis. Unfortunately, the European Parliament website does not provide any information regarding the linguistic background of the parliamentary members but only reveals their nationality. Furthermore, *Europarl* always provides the name of a speaker, but it is not

consistently tagged with the speakers' nationality. At this stage of the project, the first version of *CoStEP* did not provide the possibility to distinguish between original texts and translated material. The solution to this problem was then to manually retrieve the speakers' names from the English speakers list available on the website of the European Parliament[58], and then to update the speakers' information in the database of *CoStEP*. With this approach, it was possible to retrieve the English originals and their equivalent parallel sentences in German. The final number of parallel sentences of the data sample used for the study is 1.416 (730 in English and 686 in German)[59]: 208 in English and 204 in German for *Parliament*, 197 in English and 185 in German for *Council*, 194 in English and 193 in German for *Committee*, and 131 in English and 104 in German for *people*. The difference in the number of the English and German instances is due to wrong alignments that can occur. In these cases, the German wrong alignments were excluded from the analysis.

5.3.2 Data annotation

The annotation of the data sample was done manually and separately for English and German. It includes factors that are likely to influence article use and concerns the type of article, the noun modification, the syntactic function, and the subject-verb agreement, or concord.

The annotation for the article considers the distinction between definite article, indefinite article, and no article. Sometimes other determiners can occur instead of an article, such as demonstratives, possessives, or numeric elements; these were annotated as 'other'. The noun modification provides information about the pre- and postmodification of the analysed noun. The premodification can be a singular or a plural common noun (*e.g. the Employment Committee, the Fisheries Committee*), an adjective (*e.g. retired people, der Provisorische Legislative Rat*), a proper noun (*e.g. the Turin Council*), an acronym (*e.g. the REX Committee*), a number (*e.g. the two people, eine Million Menschen*), or a combination of them (*e.g. the Social Affairs Council*, *the Budgetary Control Committee*, *the Florence European Council*, *the Economic and Monetary Affairs Committee*). For the German part of the data, it is

[58]Online at <http://www.europarl.europa.eu/ep-live/en/plenary/search-by-organ?legislature=-1&country=GB&group=&type_organ=all>

[59] Note that all cases of *Parliament*, *Council*, and *Committee* were retrieved with upper-case, and that all cases of *people*, on the contrary, were retrieved with lower-case.

further specified if the noun is a compound, for instance *der Energieausschuss*, which is a two-element compound with *Ausschuss* as its head. The postmodification exclusively consists of prepositional phrases (*e.g. the Committee on the Environment, Public Health and Consumer Protection, der Ausschuss für Haushaltskontrolle*). The syntactic function mainly refers to the subject. Both active and passive subjects are annotated as 'subject'. A distinction was made between coordinate subject (*e.g. the Commission and the Council*) and when the collective is not the head of the subject (*e.g. Members of the European Parliament*), but they were eventually considered as 'subject'[60]. The rest (*e.g. Mr Wynn is proposing to Parliament* or *Therefore, I ask the Commission and the Council*) was considered 'not subject'. The subject and verb agreement is useful to determine if the verb following a collective noun with subject function is used in its singular or plural meaning. The morphology of English verbs is poorer than that of German (König and Gast 2009: 68-71); therefore, it is impossible to establish if the verb in English is singular or plural, for instance with modal verbs like *can*, *may* or *might*, with the future tense *will*, and with a past simple tense like *took* or *gave*. As concord is always plural with a coordinate subject, such instances were not considered in the analysis. Furthermore, when the analysed noun was not the head of the subject, the subject-verb agreement was not considered, because in this case it is not the element the verb refers to. It is however possible to establish the concord through a relative clause, for instance "the Staff Committee who advertise what they claim…" or "Abschließend möchte ich alle die Menschen, die ihren (sic) Zigaretten auf dem Schwarzmarkt kaufen".

An English sentence may also have no close translational equivalence to the German parallel sentence. In these cases, the German instance was not taken into consideration, as shown in the following examples, where the English word *people* is aligned with *Entschädigung* (i.e. *compensation*) in (1) and with *Freizügigkeit* (i.e. *mobility*) in (2):

(1)
(a) The Commissioner says that Ø people can claim […].
(b) Der Herr Kommissar sagt, dass Ø Entschädigung beantragt werden kann [...]. (*CoStEP*, 1997-05-13.xml)

[60] This differentiation might be problematic for the analysis, as the function of the collectives is different in the two contexts, but this decision is explained by the preference to keep the annotation as simple as possible.

(2)

(a) […] as a hindrance to the free movement of Ø people in the European Union.

(b) [...] als Einschränkung der Freizügigkeit innerhalb der EU betrachtet werden kann. (*CoStEP*, 1997-07-15.xml)

Finally, the data also contain information about the speakers who produced the original English instances. As mentioned above, a list of English native speakers was taken from the European Parliament website. Every speaker was then checked manually to determine if the member of Parliament was British or Irish. All this information was added and combined with the retrieved sentences.

5.4 Results and analysis

5.4.1 Parliament

The first collective noun to be discussed in the present analysis is *Parliament*. In German, this word is translated with either *Parlament* or *Haus*. Figure 5.1 shows that English and German are similar in some aspects and differ in others. Firstly, there are cases in both languages where other determiners, like possessives or demonstratives, are used instead of articles. Secondly, the indefinite article does not occur; more specifically, it is used only one time in English, shown in (3), and never occurs in German. This can be explained by the fact that there is only one Parliament to refer to and, therefore, speakers give specificity to it. Additionally, both languages do not show any postmodification.

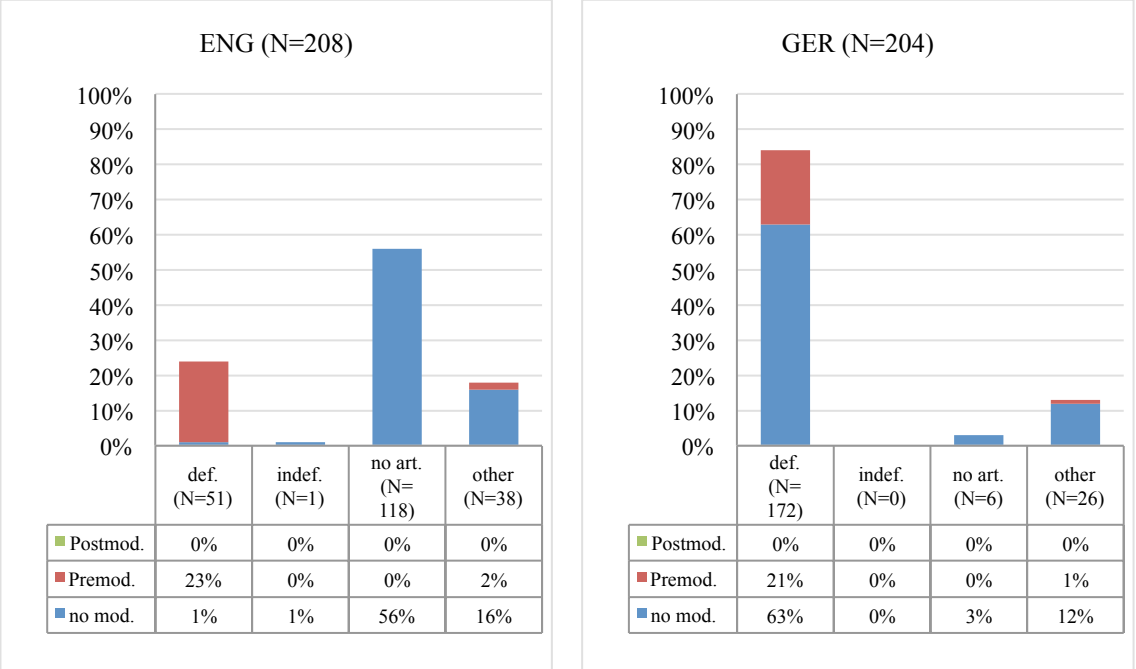

Figure 5.1: Comparison between English and German of article distribution for *Parliament*.

Turning now to the differences between English and German, one can see that these are more striking. English shows more article variability; however, article omission is strongly preferred to the use of the definite article. Furthermore, it is worth noting that the majority of nouns used with the definite article are the ones that are premodified by an adjective; two examples are given in (4) and (5). The nouns where an article is omitted, instead, have neither premodification nor postmodification, as shown in (6).

(3) We adopted, as a Parliament, a resolution in 1994 calling for the establishment of a delegation with the indigenous peoples. (*CoStEP*, 1997-02-20.xml)
(4) I have no doubt that the European Parliament will be making its views known to the heads of state before […]. (*CoStEP*, 1998-06-18.xml)
(5) The European Parliament should use this debate to make a statement and to send a message […]. (*CoStEP*, 1997-04-09.xml)
(6) I urge Ø Parliament to support the Fontain report. (*CoStEP*, 1998-07-01.xml)

On the other hand, German shows an explicit preference for article use. The results, in fact, show that the definite article occurs in more than 80% of the cases. Contrary to English, most of the cases have no premodifying elements.

5.4.2 Council

The English collective *Council* does not show any translation variety: there is only one version, namely *Rat*. Figure 5.2 shows that, in both languages, this noun is not used with other determiners. (7) is the only exception in the English data, which occurs with a possessive pronoun and with an adjective as a premodifier. Moreover, it is apparent from Figure 5.2 that both languages do not use the indefinite article, which follows the same pattern that was found for *Parliament*. In the European Parliament this institution might also be unique, therefore, it expresses something very specific.

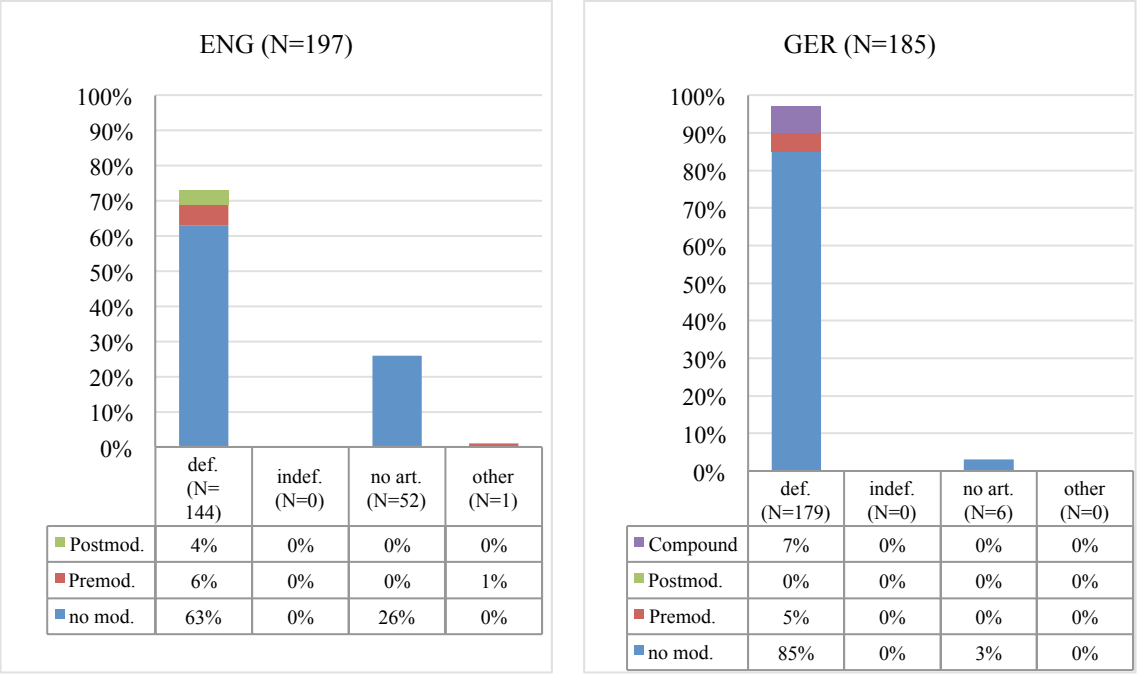

Figure 5.2: Comparison between English and German of article distribution for *Council*.

A significant difference is found with regard to article omission. German clearly favours article use. Likewise, the English parallel instances mainly show the use of the definite article; however, in contrast to German, English shows variability with article omission. Once again, as shown in (8) and (9), the cases with no article are not modified by other elements, as opposed to the ones with the definite article which, in some cases, are both premodified and postmodified, as in (10) and (11), respectively.

(7) After all, it took our own European Council six years to work out how it would have an internal electricity market here. (*CoStEP*, 1997-09-16.xml)
(8) The basic regulation only requires consultation of the industry before the Commission presents its proposals to Ø Council. (*CoStEP*, 1998-10-06.xml)

(9) I ask Ø Council to take on board those issues in the Industry Council on 7 May. (*CoStEP*, 1998-04-29.xml)
(10) […] this is actually the European Parliament's opportunity to put forward its views to the European Council which meets in Florence […]. (*CoStEP*, 1996-06-18.xml)
(11) I hope that the Council of Ministers will reconsider their position on this. (*CoStEP*, 1996-11-11.xml)

A final consideration relates to the German language. As mentioned before, instead of using premodifying units, the noun in German is combined with other elements, e.g. *Sicherheitsrat*, *Ministerrat*, and *Europarat*. These are translated in English with both premodification, i.e. Security Council, and postmodification, i.e. *Council of Ministers* and *Council of Europe*. This explains the presence of compounds in the analysis. Hence, contrary to the previous collective noun, *Rat* is also united with other words.

5.4.3 Committee

The collective *Committee* is consistently translated into German with the word *Ausschuss*. Figure 5.3 shows that there are a number of similarities between English and German.

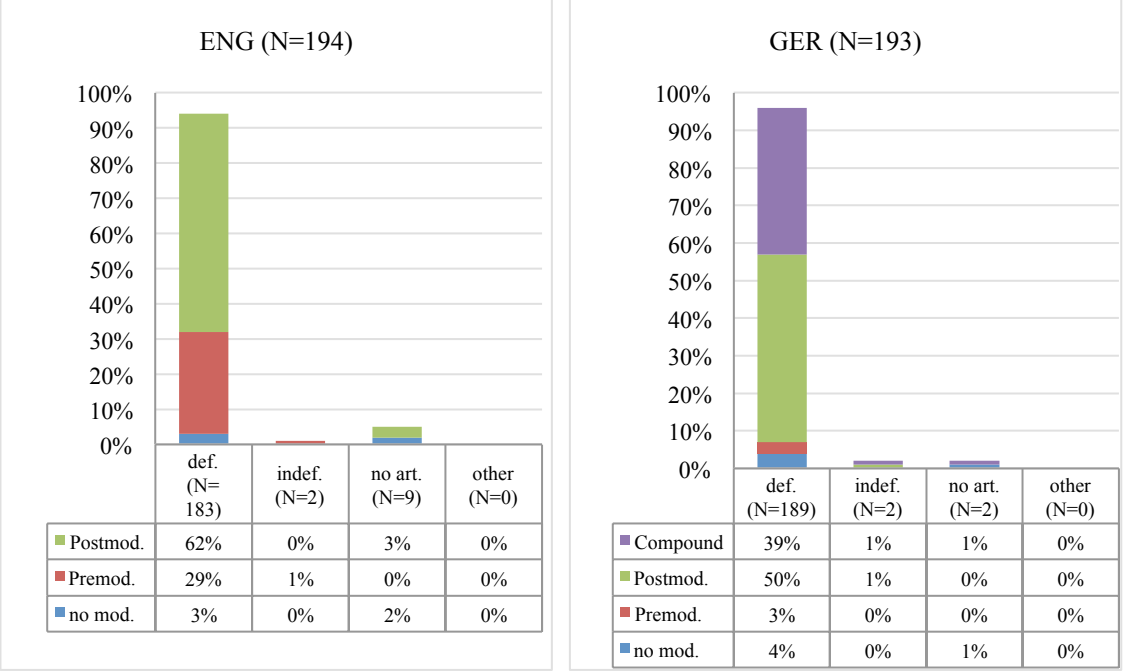

Figure 5.3: Comparison between English and German of article distribution for *Committee*.

In both languages, no other determiner is used and the indefinite article is the one with the lowest frequency. An English example is given in (12). Similarly, the presence of bare NPs is very low. (13) and (14) are two examples. On the question of the definite article, the graphs show that it is highly preferred, exceeding 90% in both datasets.

(12) If we are to continue to be an effective Budgetary Control Committee, this must continue in an unfettered way. (*CoStEP*, 1999-05-03.xml)
(13) […] which was severely lacking in the first draft presented to Ø Committee. (*CoStEP*, 2007-05-24.xml)
(14) I voted for this report, which is more balanced than the text put to Ø Committee. (*CoStEP*, 2007-05-24.xml

The most surprising aspect of the data is the fact that this collective noun, contrary to the others, rarely appears without modification. The following are a few examples taken from the English part of the data: *Committee on Petitions*, *Committee on Women's Rights and Equal Opportunities*, *Committee on the Rules of Procedures*, *Committee on Agriculture and Rural Development*, *Committee on Transport*, *Committee on the Environment,* and *Public Health and Consumer Protection*. These cases appear with postmodifiers, which are all prepositional phrases. Closer inspection of the graphs reveals that the postmodification is similar in both languages, while the premodification behaves differently. In English, it is lower than the postmodification, whereas in German its level is close to zero, but compounds are extensively used instead, such as *Fischereiausschuß*, *Haushaltsausschuß*, *Petitionsausschuß*, *Verkehrsausschuß*, *Regionalausschuß*, *Währungsaussschuß*, *Energieausschuß*, *Entwicklungsausschuß*, and *Geschäftsordnungsausschuß*. This is the most striking observation to emerge from the data comparison between English and German.

5.4.4 People

The fourth and final collective noun is *people*. It is worth noting that, in German, there are various translations for this word, which are the following: *Behinderte(n)*, *Bevölkerung*, *Bewohner*, *Bürger*, *Einwohner*, *Empfänger*, *Fischer*, *Leute*, *Menschen*, *Mitarbeiter*, *Mitglieder*, *Personen*, *Reisende(n)*, *Verbraucher*, and *Volk*. In other

words, in German there is more lexical variability. Thus, there is a tendency to be more specific and to refer to particular categories of *people*.

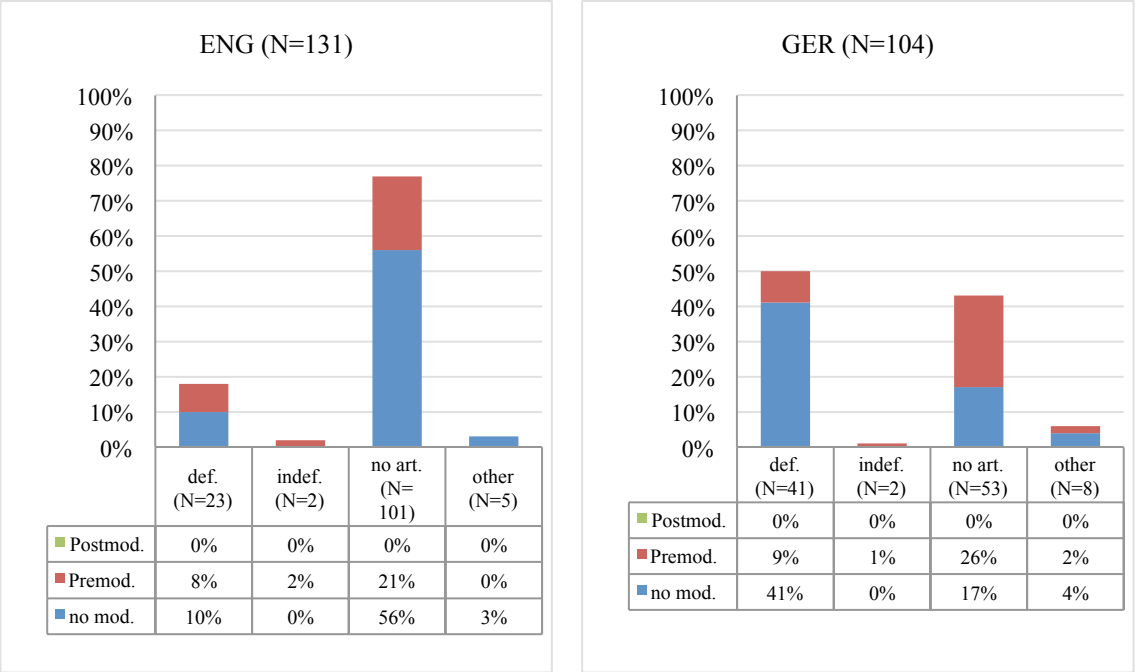

Figure 5.4: Comparison between English and German of article distribution for *people*.

This noun is different from the previous collectives because it is the only one that does not refer to or represent a legislative body of the European Parliament. As a matter of fact, the English noun *people* and its German equivalents differ from the other collectives in a number of respects. Firstly, as shown in Figure 5.4, this noun is never postmodified, but it is rather premodified, as in (15) and (16). Secondly, it presents a higher variability between the definite article and article omission, two examples are given in (17) and (18). Finally, it is clear from these graphs that, overall, this noun in the data mainly appears without modification. Nevertheless, there are also similarities, for instance other determiners are rarely used and the indefinite article exclusively occurs with elements that premodify the noun; (19) is an example from the English dataset.

(15) […] we got the security people to ask them to leave. (*CoStEP*, 1996-11-11.xml)
(16) This is a type of remortgage scheme which promised to Ø retired people the opportunity of remortgaging their property and obtaining […]. (*CoStEP*, 1997-02-19.xml)

(17) […] the Council of Ministers representing the Member States and us, in the European Parliament, representing the people, on a truly equal footing. (*CoStEP*, 1997-01-15.xml)
(18) Ø People must have access to information in their own language. (*CoStEP*, 1996-06-20.xml)
(19) I know the BSE problem is indeed a very grave one and that an estimated 15 people have died from the associated disease of CJD. (*CoStEP*, 1997-02-20.xml)

With respect to the comparison between the two languages, one can note that English considerably differs from German with regard to the higher use of bare NPs, which is highly preferred to the definite article. As shown in (20)a, (21)a, and (22)a, these cases refer to *people* with generic reference. By contrast, German exhibits high variability between the definite article and article omission, but with a preference for the definite article, as shown in the parallel sentences of the previous examples in (20)b, (21)b, and (22)b.

(20)
(a) The failure to update and review it, to protect Ø people from the effects of radiation rather than to promote nuclear power, is extraordinary. (*CoStEP*, 1996-12-11.xml)
(b) […] daß dieser Vertrag nicht auf den neuesten Stand gebracht und überprüft wird und die Menschen vor den radioaktiven Strahlungen geschützt werden, anstatt die Kernkraft zu fördern. (*CoStEP*, 1996-12-11.xml)

(21)
(a) Every Member State has its own sneaky little tricks for making sure that Ø people do not find it too easy to go to other countries and practice their professions. (*CoStEP*, 1997-06-09.xml)
(b) Jeder Mitgliedstaat hat seine eigenen kleinen hinterhältigen Tricks, um dafür zu sorgen, daß es dem Bürger nur nicht zu leicht gemacht wird, seinen Beruf in einem anderen Staat der EU auszuüben. (*CoStEP*, 1997-06-09.xml)

(22)
(a) This register would simply give information to those seeking to employ Ø people in the form of a 'yes' or 'no' about a previous conviction. (*CoStEP*, 1996-09-18.xml)
(b) Dieses Register würde denjenigen, die Leute einstellen wollen, lediglich in Form von „ja" oder „nein" Auskunft über Vorstrafen geben. (*CoStEP*, 1996-09-18.xml)

However, the data show that German, similar to English, expresses genericness by omitting the article, too, as shown in the following parallel examples.

(23)

(a) One of the key issues is the question of whether doctors and medical personnel will be free to diagnose and treat Ø people without a royalty payment to the patent-holder. (*CoStEP*, 1997-07-15.xml)

(b) Eine der Schlüsselfragen ist, ob Ärzte und medizinisches Personal Diagnosen stellen und Ø Menschen behandeln dürfen, ohne dem Patentinhaber Gebühren zahlen zu müssen. (*CoStEP*, 1997-07-15.xml)

(24)

(a) […] the fact is that Ø people leaving the Community would still be entitled to buy duty-free. (*CoStEP*, 1997-07-16.xml)

(b) [...] wurde im Hinblick auf die Randlage bereits erwähnt, daß Ø Reisende, die die Gemeinschaft verlassen, weiterhin zollfrei einkaufen dürften. (*CoStEP*, 1997-07-16.xml)

It is worth noting that there are more German bare cases in premodified contexts, with an adjective, as in (25), or numeric elements, as in (26).

(25) Ø Behinderte Menschen sind in allen Bereichen ihres täglichen Lebens direkter oder indirekter Diskriminierung ausgesetzt. (*CoStEP*, 1996-12-12.xml)
(Ø Disabled people experience direct and indirect discrimination in all areas of their daily lives.)

(26) Kürzlich starben 15 Menschen in einem Mitgliedstaat an Lebensmittelvergiftung, nachdem sie verseuchte Lebensmittel aus einem kleinen Großhandelsunternehmen gegessen hatten. (*CoStEP*, 1997-02-20.xml)
(Recently 15 people died in one Member State from food poisoning as a result of eating contaminated food from one small wholesale establishment.)

However, the results indicate that the premodification is more or less equal in both languages, but, as previously mentioned, this collective is mainly used without modification.

Taken together, these results suggest that, with the collective nouns in question, English and German behave similarly. However, it is clear that bare NPs are more likely to occur in English, whereas German tends more toward article use. The next section describes the evaluation regarding the relation between the syntactic function and the subject-verb agreement.

5.4.5 Syntactic function and concord

The annotation for the syntactic function distinguishes between [+subject] and [–subject]. As previously mentioned, the pattern [+subject] also contains the cases where the element is a *coordinate subject* or *part of the subject*.

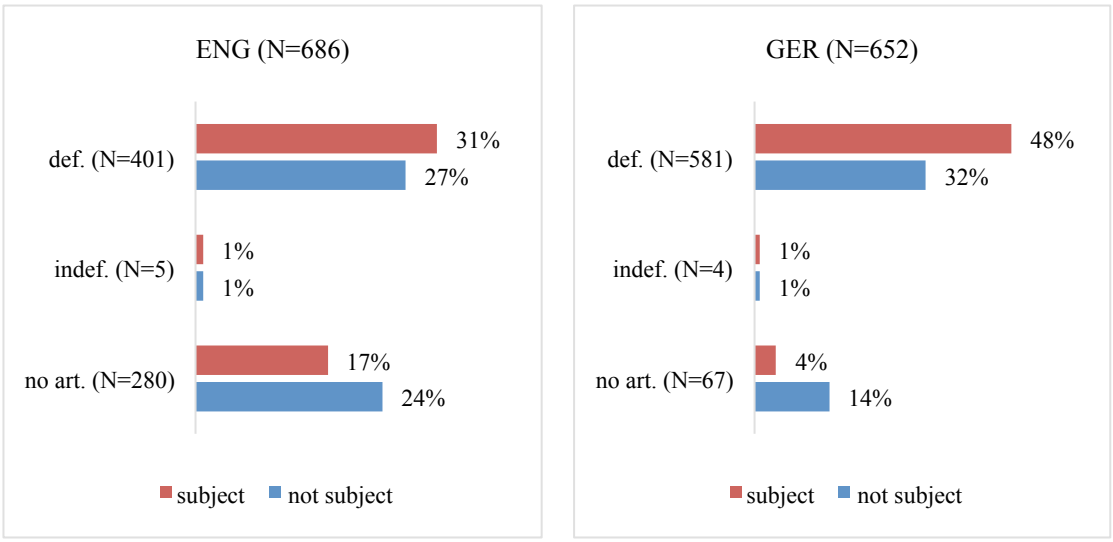

Figure 5.5: Comparison of distribution of syntactic function among articles between English and German.

Figure 5.5 shows the distribution of the syntactic function among articles[61] and compares English and German. As seen in the previous analysis, the indefinite article is rarely used with the collectives taken into consideration. What is more interesting regards the cases occurring with the definite article and those omitting an article. From the graphs, it can be seen that, in both languages, the definite article is more used in subject position; by contrast, bare NPs occur more often in non-subject position. Furthermore, the data show that this distinction is more pronounced in German than in English. Additionally, the results point to a tendency in English of more variability between the use of the definite article and article omission. On the other hand, in German, the definite article is preferred to the omission of an article. However, it is noteworthy that for this investigation it was not possible to analyse if the collective has an anaphoric reference, i.e. if it was already mentioned before. The data do not show the sentences that precede the clause in which the collective is used.

[61] Therefore, this analysis does not include the cases that occur with another determiner.

In other words, it is not possible to know the context. This information might be relevant to better understand the variability, particularly in English.

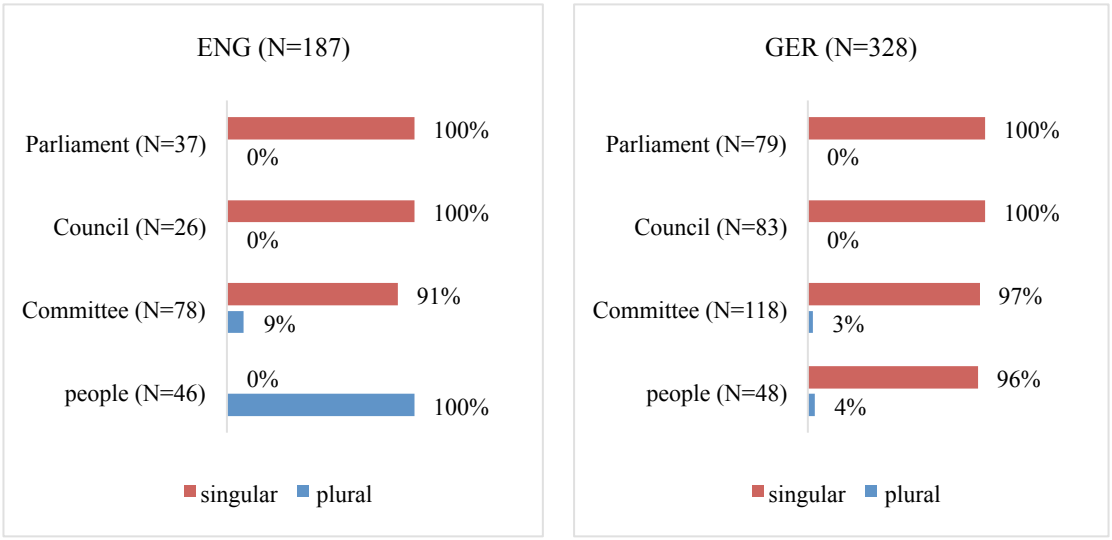

Figure 5.6: Comparison of subject-verb agreement between English and German.

Another interesting point for the analysis is to see whether the number of the verb following the noun in subject position influences article use. As seen in the introduction section of the current chapter, this issue has been insufficiently researched. The results of this analysis are presented in Figure 5.6 above.[62] The data show that in both languages the verbs following the nouns *Parliament* and *Council* are always used in their singular form. Furthermore, the graphs indicate that there is slight variability with the collective *Committee*, whose verb, however, is also mainly used in the singular in both English and German. (27) and (28) are examples of *Committee* with singular concord, while in (29) and (30) the subject-verb agreement is plural.

(27) […] in exactly the same way on the floor of this House before the Committee has completed its deliberations. (*CoStEP*, 1998-03-09.xml)
(28) Obviously the Committee agrees with the Commission assessment that it would be unrealistic to allocate the European Environmental Agency […]. (*CoStEP*, 1998-02-18.xml)
(29) […] I have been represented by the Staff Committee who advertise what they claim to be the price of Tomlinson. (*CoStEP*, 1997-06-25.xml)

[62] Note that only the cases whose subject-verb agreement could be identified were taken into consideration.

(30) […] which resulted in the recommendations which the Temporary Committee have been seeing through. (*CoStEP*, 1997-11-18.xml)

These results then confirm what previous studies (e.g. Depraetere 2006, Levin 2006, Collin Online 2017) have stated. By contrast, the noun *people* is the only one that presents a difference between the two languages. In English, this collective is constantly used in the plural. This is also attested by Jespersen (1949: 94), Kruisinga (1960: 62) and Depraetere (2003: 114), who affirm that collectives denoting living beings are used with the plural construction. By contrast, in German, a small percentage of the verbs are singular. This is explained by the fact that German uses different translations for the English word *people*. Therefore, there are some nouns which require a plural verb (e.g. *Leute* and *Menschen*), similar to English, and there are other nouns that are used with a verb in singular (e.g. *Volk* and *Bevölkerung*). Finally, the high use of the verb in plural with the noun *people* obviously explains the low frequency of the indefinite article. Overall, these findings suggest that the agreement between the subject and the verb is influenced by the meaning of the collective noun and that does not affect the use of articles.

5.4.6 British English vs. Irish English

As mentioned before, the sentences of the English data sample were originally produced by native speakers. Therefore, it is important to remember that, in the European Parliament, there are speakers from both the United Kingdom and Ireland. As already discussed in section 2.8 and in the introduction of the current chapter, it is known that IrE uses the definite article more frequently than BrE (Hickey 2007: 251, Corrigan 2010: 52, Kallen 2013: 122). Thus, it is interesting to investigate if the differences between the two varieties are confirmed, or if they show similarities. All in all, there are 88 English speakers: 17 from Ireland, 67 from the United Kingdom, and 4 are unknown. Figure 5.7 shows the distribution of the definite article, indefinite article, and cases that omit an article among the two varieties of English.[63]

[63] Note that the analysis only excludes the cases where another determiner occurs and the English native speakers, whose origin is unknown.

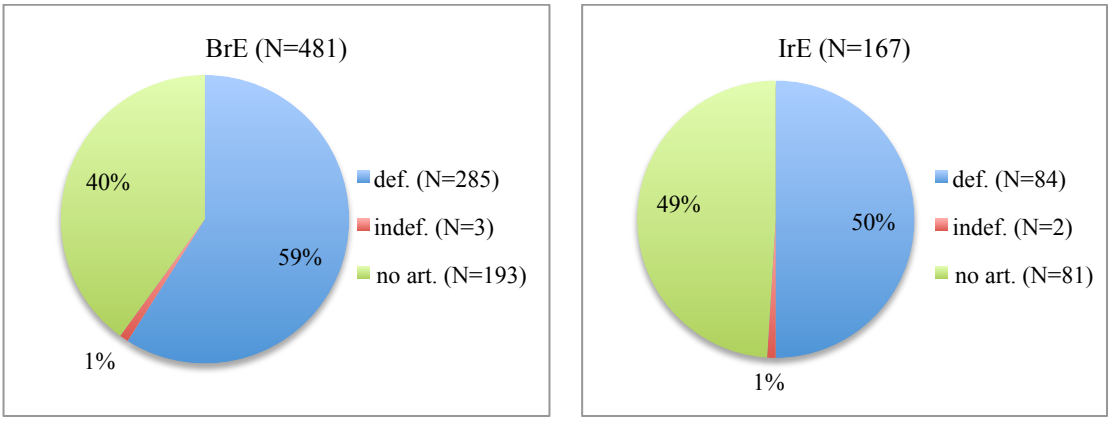

Figure 5.7: Comparison of article distribution between British and Irish speakers.

Firstly, the pie charts confirm that the speakers hardly use the indefinite article, which occurs only for 1% of the data in both language varieties. Secondly, in both BrE and IrE, the definite article is used slightly more often than the omitted article. However, this distinction is more evident in BrE (i.e. 59% vs. 40%). By contrast, in IrE, the distribution of the definite article and the cases occurring with no article is more equal (i.e. 50% vs. 49%). It is worth noting that the number of speakers is not equally distributed among the two English varieties. Therefore, a chi-square test was used to test whether this difference was significant. The results are presented in Table 5.1.[64]

	χ^2	*p*
BrE vs. IrE	3.23	0.06

Table 5.1: Significance of article use between BrE and IrE.

The table shows that the difference is slightly above the significance level (i.e. $p < 0.05$). This data thus do not strongly confirm the fact that IrE uses articles extensively. However, as shown in section 4.2, transcribers tend to favour traditional forms and to be more conservative with regard to grammar rules. Example (18) discussed in the same chapter showed that the expression *the Parliament*, uttered by an Irish speaker, was modified and changed into *Parliament* because the use of the definite article was considered ungrammatical. It is therefore possible that the data do not present

[64] The indefinite article was combined with the definite article, due to its low frequency.

significant results because the cases of article overuse in IrE might not have been reported in the transcriptions.

5.4.7 Logistic regression analysis

Since the data come from a parallel corpus, it is necessary to do the statistical analysis separately for English and German. For the logistic regression analysis, article use is the dependent variable, and four independent variables are included: speaker origin, modification, noun type, and syntactic function.[65] The cases where a different determiner was used instead of an article (i.e. a demonstrative or a possessive) were excluded. The category 'speaker origin' simply distinguishes between 'BrE' and 'IrE'. In order to simplify the statistical model, the rest of the annotation had to be modified. Firstly, the definite and indefinite articles were combined into one category called 'article'. This allowed to have a binary dependent variable: 'article' (i.e. presence of an article) and 'no article' (i.e. absence of an article). Secondly, due to the low frequency of some premodifiers, the category called 'modification' was created to merge the factors 'premodification', 'postmodification', and 'no modification'. Finally, the factors 'coordinated subject' and 'part of the subject' were combined into one factor named 'subject'. The category of the syntactic function, therefore, distinguishes between 'subject' and 'not subject'. Table 5.2 gives the results for English.

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**Significance**
(Intercept)	-0.636	0.679	-0.936	0.349	
postmodif.	-2.414	0.737	-3.275	0.001	**
premodif.	-3.451	0.368	-9.356	< 2e-16	***
BrE	-0.072	0.264	-0.273	0.785	
Council	0.001	0.635	0.002	0.998	
Parliament	2.800	0.647	4.326	1.52e-05	***
people	4.024	0.693	5.803	6.53e-09	***
subject	-0.151	0.238	-0.635	0.525	

Table 5.2: Results of the logistic regression analysis for the English dataset.

[65] Note that in both English and German the category on the subject-verb agreement was not included in the logistic regression analysis, due to the high number of unavailable values in the annotation.

The reference level used in the statistical model is from the perspective of article use.[66] The results reveal that the nouns *Parliament* and *people* are highly significant (marked as '***'), i.e. they are very likely to occur as bare NPs. Furthermore, the presence of modifiers influences article use and shows significant results; more specifically, 'postmodification' is significant (marked as '**'), while 'premodification' is highly significant (marked as '***'). This means that when a collective noun is postmodified or premodified, it is very likely to occur with an article. Based on these results, the syntactic function does not play a role in article use with collectives; moreover, they confirm that the speaker nationality does not influence article use. In Table 5.3, the factors are sorted by odds ratio (i.e. ranked based on their factor strength): from the strongest to the weakest influencing article omission.

Factor	**Odds ratio**[67]
people	56.00
Parliament	16.44
Council	1.00
BrE	0.93
subject	0.85
postmodification	0.08
premodification	0.03

Table 5.3: Ranking of factors influencing article omission in English.

The results show that the noun *people* tends towards article omission more strongly than *Parliament*. By contrast, the factors 'postmodification' and 'premodification' are less likely to occur without an article; however, as already seen in Table 5.2, the presence of a premodifier in the NP slightly favours article use more strongly than a postmodifier. A good way to investigate the significance level of the different categories used in the first statistical model shown in Table 5.2 is to rerun the statistical model without the variable of interest and compare the fit of the two models. The results are summarized in Table 5.4.

[66] Note that the coefficients in front of the factors reveal how more/less likely is to observe the non-reference compared to reference.
[67] The odds ratio values are rounded to two decimals.

Independent variable	p value	Significance
noun type	$< 10^{-41}$	***
modification	$< 10^{-28}$	***
syntactic function	0.52	
origin	0.78	

Table 5.4: Significance level of single independent variables in English.

Comparing the first statistical model with the one in which the 'noun type' category is removed shows that the difference between the two models is highly significant (marked as '***'). This means that the 'noun type' strongly influences article use; put differently, the use of articles with collective nouns is lexeme-dependent. Similarly, the difference between the first statistical model and the one in which the category 'modification' is removed also shows highly significant results (marked as '***'). Finally, as already seen in Table 5.2, the categories 'syntactic function' and 'speaker origin' are not significant and, therefore, do not play a role in article use.

Moving on now to consider the German part of the data, the annotation had to be simplified accordingly. The category called 'article' combines the definite and indefinite article and distinguishes between two factors, namely 'article' and 'no article'. Furthermore, 'modification' includes four factors, i.e. 'no modification', 'premodification', 'postmodification', and 'compound'. Finally, the 'syntactic function' distinguishes between 'subject' and 'not subject'. Since the investigation focuses on English native speakers (i.e. English originals), the category regarding the origin of the speaker in the German translations was not included in the analysis. Similar to English, the logistic regression analysis for German has a binary dependent variable, i.e. the presence or absence of an article, but only three independent variables, i.e. 'noun type', 'modification', and 'syntactic function'. The results are shown in Table 5.5.

	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-19.557	1095.944	-0.018	0.985	
no modif.	17.014	1095.944	0.016	0.987	
postmodif.	0.013	1539.230	0.000	1.000	
premodif.	18.145	1095.944	0.017	0.986	
Council	-0.864	0.892	-0.969	0.332	
Parliament	-1.194	0.877	-1.362	0.173	
people	1.933	0.800	2.416	0.015	*
subject	-0.031	0.328	-0.095	0.924	

Table 5.5: Results of the logistic regression analysis for the German dataset.

The logistic regression analysis reveals that German differs from English; there is, in fact, only one significant factor, namely the noun *people* (marked as '*'), which tends to occur more often without an article. It is almost certain that this particular noun shows significance because, in comparison to the other collectives included in the analysis, it is the one with the highest frequency of bare NPs (i.e. 53 out of 66). In the German dataset, the use of an article is categorical and the statistical results are then influenced by the underlying frequencies. However, for sake of completeness, Table 5.6 shows the list of factors ranked by their effect size (i.e. sorted by their odds ratio), from the biggest to the smallest, i.e. from the strongest to the weakest influencing article omission.

Factor	**Odds ratio**
no modification	7.59×10^{7}
premodification	2.45×10^{7}
people	6.91
postmodification	1.01
subject	0.96
Council	0.42
Parliament	0.30

Table 5.6: Ranking of factors influencing article omission in German.

The results show that 'no modification', 'premodification', and 'people' are the factors with the biggest effect size on article omission, while, at the end of the ranking, we find 'Parliament' and 'postmodification' with the smallest effect size values.

Modification factors	**article**	**no article**
no modification	*N* 346	*N* 34
premodification	*N* 70	*N* 32
postmodification	*N* 99	*N* 0
compound	*N* 93	*N* 0

Table 5.7: Article distribution among the factors of the modification category.

At first, the two factors on the top might be surprising, but a better look at the data reveals that 'no modification' and 'premodification' show a strong preference towards

article omission because these are compared to the other two factors found within the same factor category, i.e. 'postmodification' and 'compound'. The distribution of the articles among these four factors, reported in Table 5.7, shows that 'no modification' and 'premodification' strongly favour article omission because they are in contrast to 'postmodification' and 'compound' which never occur without an article.

Independent variable	p value	Significance
noun type	$< 10^{-19}$	***
modification	$< 10^{-05}$	***
syntactic function	0.92	

Table 5.8: Significance level of single independent variables in German.

Table 5.8 shows the significance level of the independent variables, analysed separately. This was possible by removing the independent variables from the first statistical model and comparing the results of the reduced models to the original one (i.e. the model including all the independent variables). Similar to English, the categories of 'noun type' and 'modification' show significant values, while the syntactic function does not play any role in article use with collective nouns. However, as mentioned before, German differs from English because it uses articles much more consistently, and the frequencies of bare NPs are notably low. These results therefore need to be interpreted with caution.

5.5 Summary

The present study compared the use of articles in English and German with a defined class of nouns, i.e. collectives. The analysed nouns were the English *Parliament*, *Council*, *Committee,* and *people* and their equivalent translations in German. The results have shown that these two languages share similarities but also exhibit differences. With *Parliament*, article use greatly differs between the two languages. German prefers the definite article, while English favours article omission and mainly uses the definite article when a noun is premodified. In comparison, *Council* presents more similarities between the two languages, with a higher use of the definite article. However, English shows variability between the definite article and article omission. *Committee* functions differently from the other nouns, because it barely appears without modification but mainly occurs with postmodification and premodification in

English, and postmodification and compounds in German. This might explain the reason why both languages favour the use of the definite article. Lastly, *people* shows differences between English and German. The former prefers the article to be omitted, the latter favours the definite article. However, it is important to note that in German articles are amply omitted, but most cases occur with premodification. Moreover, German translates the English word *people* in various ways, showing more lexical variety and specificity regarding the category of people that the context refers to. In summary, the most particular case is *Parliament*. This collective noun can also refer to an institution and, according to Christophersen (1939: 182), "[l]egislative bodies when regarded as permanent are in zero-form." It could be possible that *Parliament* is used in the same way, regardless of whether it refers to the collective number of individuals or to the institution. However, this is only the case for English, because German, on the contrary, uses an article more extensively. A likely explanation is that in German it is still seen as the unique collective unit or institution of the European Parliament. Therefore, German speakers tend to refer to it with specificity. By contrast, in English, *Parliament* could be seen as a permanent group of people or institution; this in turn might make *Parliament* share similar characteristics of proper nouns, which generally do not require an article.

The examination with regard to syntactic function attested that in both English and German the definite article is mainly used with nouns in subject position, while articles are principally omitted with nouns in non-subject position. Furthermore, the subject-verb agreement is exclusively determined by the meaning of the collective noun. In addition, the distribution of articles among British and Irish speakers did not show particular dissimilarities. Both varieties present similar variability between article use and article omission, with a preference for the former. The chi-square test also confirmed that the difference is not significant and it is therefore not possible to strengthen the observation that IrE uses articles more often than BrE for the used data (e.g. Hickey 2007, Corrigan 2010, and Kallen 2013). This result may be explained by the fact that unusual expressions are generally modified by transcribers, who might have in turn preferred to exclude possible cases of article overuse in IrE. However, in order to have a better perspective, an equal number of speakers would be necessary.

Finally, the logistic regression analysis showed that, in English, the syntactic functions and the origin of the speaker had no significance in article use. It confirmed, however, that the collectives *Parliament* and *people* are more likely to occur as bare

NPs and that a collective tends to use an article when it is modified: modification appeared to be significant. The statistical analysis then reinforced the fact that article use is influenced by the noun modification and by the type of the noun. On the other hand, in German, the logistic regression analysis showed that the only collective favouring article omission is the German equivalent of *people*. However, with a very low frequency of bare NPs, caution must be applied when interpreting the results of the statistical analysis. Furthermore, it is not possible to generalize about the category of collectives. As stated by Poutsma (1904) and Christophersen (1939), articles are less regular and not easily predictable with regard to collective nouns. Overall, however, the results strengthen the idea that article use is mainly influenced by the actual meaning of the collective noun and by the modifying elements, which make the noun more (or less) specific. Finally, according to the results of the investigation, it is clear that German uses articles more frequently, while in English they are more variable. Hence, German proved to be a suitable language source for the purpose of the current project, i.e. retrieving English bare NPs via aligned German NPs occurring with an article.

In conclusion, the contrastive approach in the present study was useful to investigate English and German with regard to article use and to analyse their similarities and differences with collective nouns. Moreover, it showed the opportunities that a parallel corpus can provide for (contrastive) language investigation. In the following chapter, the analytic focus will finally change to English only, i.e. the intended purpose of the parallel corpus data retrieval all along. In particular, attention will be paid to the data-driven method adopted for the data retrieval of bare NPs in English and to the analysis of (variable) article use from a Construction Grammar view.

6 Data-driven approach: variable articles across constructions

6.1 Introduction

The great advantages provided by (parallel) corpora have been described in numerous previous studies (see e.g. Church and Mercer 1993; Greenbaum 1996; Kennedy 1998; Johansson 2002). A key point in (Parallel) Corpus Linguistics is that the evidence comes from existing data and, therefore, we do not "loo[k] at what is theoretically possible in a language, [but] we study the actual language used in naturally occurring texts" (Biber, Conrad and Reppen 1998: 1). A corpus-driven approach can thus be immensely relevant for the investigation of linguistic variation. With the vast amount of data provided by large (parallel) corpora, it is possible to examine different contexts in which article use is variable. Furthermore, the use of corpus data is considered to be of importance in the field of Construction Grammar, because "corpora can provide empirical support to intuitions regarding the nature of constructions and the number of constructions in the constructional inventory" (Trousdale and Gisborne 2008: 71). The combination of a data-driven method with a constructional perspective, as adopted in this study, results in a bottom-up analysis and allows for a deeper insight into variable English article use.

The focus of this chapter is on the methodology adopted for the retrieval of contexts in which articles are variable in English, the descriptive analysis thereof, and a proposed model accounting for article use/variability in a Construction Grammar framework. The following section illustrates the data-driven selection of bare NP corpus data, followed by a detailed analysis thereof (section 6.3). Section 6.4 will then look in detail at a particular case-study discovered during this data-driven investigation, before moving on to the presentation of a Construction Grammar model of (variable) article use based on empirical evidence from the investigated corpus data (section 6.5). The summary of the findings is given in section 6.6.

6.2 The retrieval process of the bare NPs dataset

The data-driven approach proceeded in several steps. At first, a sample of 500 random sentences was retrieved where the German version contained noun phrases with an article (either definite or indefinite), aligned with parallel English noun phrases without an article (e.g. *Derzeit unterscheiden sich die Normen – Certainly, at the*

moment, Ø standards vary; *Die Umverteilung muß gerecht sein – Ø Re-distribution must ensure fairness*).

In this initial phase, common errors usually relate to part-of-speech tags, dependencies (i.e. parser errors), word alignments, correlations, translation imprecisions, and German compounds.[68] In order to improve the corpus parser and to refine the succeeding data retrieval, a preliminary evaluation of the aligned elements and their parse-dependencies was needed. The evaluation was done manually and the first sample was used as a basis to identify and better define the filters necessary to avoid false positives (i.e. noun phrases that did not contain an article but were not bare NPs). The first restriction regarded all NPs in German aligned with an NP in English preceded by a possessive pronoun (e.g. *my, your, his, her*) or a demonstrative determiner (e.g. *this, that, these, those*). These elements fill the same slot in which an article may occur and can therefore not be considered bare NPs. Likewise, English nouns preceded by the numeral *one*, as in "only one part of […]", cannot be counted as articleless, as *one* can be realized as the indefinite article *a/an*. Therefore, these cases were also part of the filters. Second, Saxon genitives in the English parallel sentences were identified as another filter for data retrieval. In a Saxon genitive construction such as (*the) workers' rights* or *the Commission's proposal*, the article generally relates to the second element, namely the element that is possessed (i.e. *rights* and *proposal*), and not the first one, namely the element that possesses (i.e. *workers* and *Commission*). The reason why these cases could not be included was due to a parser error, which, at this stage, occasionally marked the article as dependent on the first element and not the second one. Even though these cases could have been filtered out manually at a later stage, they were automatically excluded from the beginning in order to have as clean data as possible.

A further common problem concerned English noun premodifiers. English translation equivalents of German nouns (especially compounds) often correspond to a sequence of nouns that are dependent on each other, and the article always refers to the head noun. In other words, only the head noun, which all other components are dependent on, is of importance. Unfortunately, at times, word alignments were wrong, making the analysis of article use impossible. Example (1) is illustrative of an

[68] It is known that German frequently uses compound nouns. These can sometimes be problematic for word alignments and the retrieval of parallel data. A 1:1 correspondence can fail because English generally translates German compounds as a sequence of nouns.

alignment error: the German noun *das Geld* is mistakenly aligned in English with *consumers* instead of *money*.

(1)

(a) Mit einem Streich schützt sie die Umwelt, bewahrt sie Ressourcen, führt sie zu einem wirtschaftlichen Aufschwung und spart gleichzeitig das Geld der Verbraucher. (*CoStEP* 1996-04-15.xml)
(b) At one stroke it protects the environment, conserves resources, boosts the economy and saves consumers money at the same time. (*CoStEP* 1996-04-15.xml)

Finally, coordinated structures with *and* are sometimes ambiguous, and the question arises whether the noun following the conjunction is within the scope of the article or not. According to Quirk et al. (1985: 960), "[w]hen coordinated heads are preceded by a determiner, the usual interpretation is that the determiner applies to each of the conjoins." Therefore, expressions like *a knife and fork* or *the head and shoulders* correspond to *a knife and a fork* and *the head and the shoulders*. However, as shown in (2), article use in coordinated noun phrases is variable. These cases were therefore not filtered out.

(2)

(a) […] must not be won at the cost of the hopes, the needs and the aspirations of so many of our people. (*CoStEP* 1996-10-23.xml)
(b) […] an approach which is sensitive to the needs and Ø aspirations of the homeless, which takes account of […]. (*CoStEP* 1997-05-28.xml)

After the evaluation of the first sample and the identification of possible restrictions, a second sample of 500 randomized parallel concordances was retrieved. The filters evaluated in the first sample were applied here and another evaluation was done manually in order to check the validity and reliability of the second retrieval process. From this second sample, 200 refined parallel occurrences were randomly selected, manually checked, and used to identify repeatedly occurring parsing patterns in which an article could vary. The latter were manually identified based on the dependency relations between the elements of an NP; they allow for a more structured approach to investigate potential article variation contexts. The following are the identified patterns: SUBJECT PHRASE, PASSIVE SUBJECT PHRASE, PREDICATIVE CLAUSE, OBJECT PHRASE, NON-FINITE PHRASE, PREPOSITIONAL PHRASE (FOLLOWING A NOUN), PREPOSITIONAL PHRASE (FOLLOWING A VERB), and COORDINATED PHRASE. These 8

parsing patterns were then used as a filter to evenly retrieve more data from the whole corpus,[69] i.e. a grand total of 1,200 parallel sentences per language.[70]

Again, the sample contained false positives that had to be manually removed. In addition to the previously mentioned problems (i.e. the use of a demonstrative, possessive, or numeral instead of an article), more false positives were filtered out. These included part-of-speech annotation errors in which the item in question was not a noun but a pronoun (e.g. *that is certainly someone*), or an adjective (e.g. *a solution which is WTO-compliant*), or a verb (e.g. *the country plunges into*). Moreover, in some cases, the item in question was part of a more complex noun phrase (e.g. *the European Renewable Energies Export Council*) and noun sequences (e.g. *the diplomatic and inspections options*) that were preceded by an article. Furthermore, the retrieval generated instances belonging to other patterns. When possible, these were manually corrected and kept in the dataset. Additionally, the retrieval yielded incomplete instances and instances including dates (e.g. *11 September, 9 April*); these were also excluded. Finally, the COORDINATED PHRASE was removed from the investigation due to parsing inconsistencies and difficulties in the later stages of the analysis as an NP from a coordinated structure can simultaneously be part of one of the other structures (i.e. not independent). After the exclusion of the above cases, the dataset contained a total of 851 instances with bare NPs. Before proceeding to the descriptive analysis of the bare NPs dataset, the parsing patterns are explained in more detail.[71]

Example (3) refers to the PREPOSITIONAL PHRASE PATTERN (FOLLOWING A NOUN), while (4) refers to the PREPOSITIONAL PHRASE PATTERN (FOLLOWING A VERB).

(3) In this area Bavaria is quite definitely in the front line of the fight against international crime on behalf of the entire […] (*CoStEP* 1996-06-20.xml)
(4) […] we can see the problems that arise with reprivatization, just as they do with privatization. (*CoStEP* 1996-10-23.xml)

[69] In order to avoid bias towards more frequent patterns, 150 instances per pattern were retrieved. For a frequency list of the pattern distribution over the whole corpus, see Appendix A.
[70] Note that the English part of the data only includes instances uttered by native speakers.
[71] For the visualization of the retrieval patterns, noun phrases are referred to using the original PoS tags, namely NN, NNS and NP, which stand for common noun in singular, common noun in plural, and proper noun, respectively.

The corresponding visualizations are illustrated together in Figure 6.1 and show that, in a prepositional phrase the focus is on the use of articles in front of the noun phrase that follows the preposition.[72]

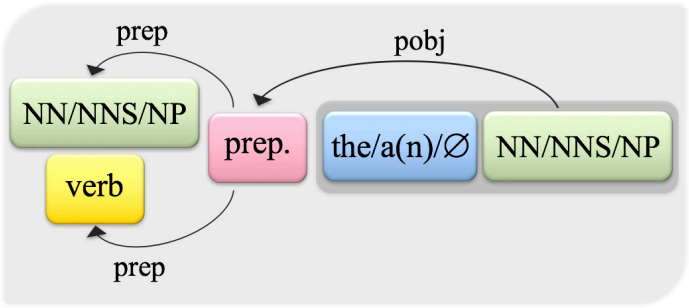

Figure 6.1: Parse dependencies for the prepositional phrase patterns.

The only difference between the two patterns is that in the PREPOSITIONAL PHRASE PATTERN (FOLLOWING A NOUN) the preposition follows a noun, while in the PREPOSITIONAL PHRASE PATTERN (FOLLOWING A VERB) the preposition is attached to a verb.

Both the OBJECT PHRASE PATTERN, as in (5), and the NON-FINITE PHRASE PATTERN, as in (6), refer to a direct object.

(5) Thus we need to establish a coherent European tourism policy which adds value above and… (*CoStEP* 1996-10-24.xml)
(6) […] it is equally important to strengthen public-service production and distribution. (*CoStEP* 1996-10-21.xml)

Figure 6.2 shows the graphic visualizations of the OBJECT PHRASE PATTERN and the NON-FINITE PHRASE PATTERN. The focal point is the possibility to include an article in the NP that follows the verb.

[72] The direction of the arrow underlines the dependency relation between two elements. For instance, if the arrow points from A to B, this means that A depends on B.

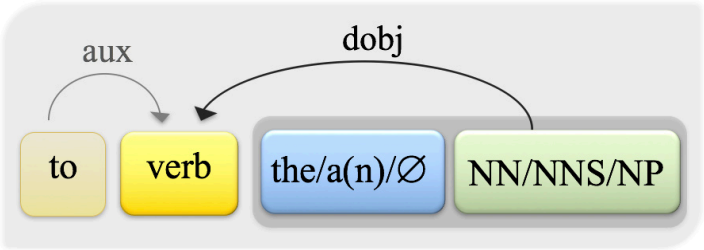

Figure 6.2: Parse dependencies for the object phrase and non-finite phrase patterns.

The only difference between the two chunks is that the verb in the OBJECT PHRASE PATTERN is finite, while it is non-finite in the NON-FINITE PHRASE PATTERN and accompanied by *to* (marked in lighter shading).

Sentences (7) and (8) are examples of the SUBJECT PHRASE PATTERN and the PASSIVE SUBJECT PHRASE PATTERN, respectively.

(7) Citizens need to know there is someone, a real human being, who will take their side is bureaucracy threatens to ignore the [...] (*CoStEP* 1996-06-20.xml)
(8) There is no freedom of opinion or of the press; on the contrary, journalists are being prosecuted too. (*CoStEP* 2006-11-16.xml)

The visualization of the dependencies between the elements of the SUBJECT PHRASE PATTERN and the PASSIVE SUBJECT PHRASE PATTERN are given in Figure 6.3. In both chunks, an article can occur in front of the noun, which in turn depends on the verb that follows.

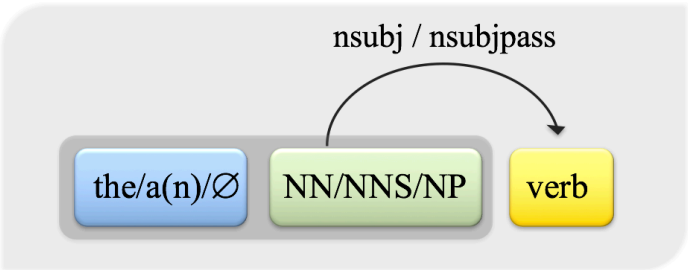

Figure 6.3: Parse dependencies for the subject phrase and passive subject phrase patterns.

The difference between these patterns regards the dependency type between the noun within the NP and the verb, i.e. passive or active subject.

Finally, an example of the PREDICATIVE CLAUSE PATTERN is provided in (9). In this pattern, two NPs and two dependencies are involved, as shown in Figure 6.4.

(9) Export refund are economic nonsense. (*CoStEP* 1997-10-24.xml)

One dependency relates to a subject relation (coded as *nsubj*) between the nouns, while the other is coded as *cop* (i.e. *copula*) and refers to a copula verb, which precedes the second NP (i.e. predicate noun). In this chunk, article variability occurs in the NP that follows the copula verb.

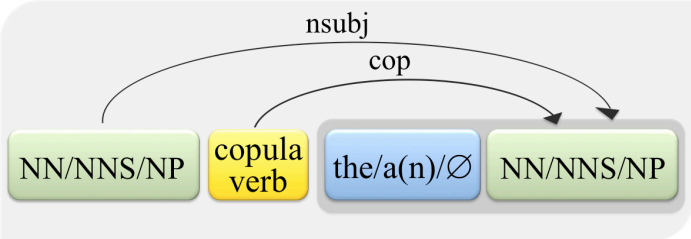

Figure 6.4: Parse dependencies for the predicative clause pattern.

It is worth noting that, in all chunks, the noun phrase can be complex; therefore, more elements (e.g. an adjective) can occur between the determiner and the following noun, as shown for instance in (3), in which the common noun *crime* is preceded by the adjective *international*. The following section will describe the dataset containing the bare NPs and will highlight the main features (e.g. what types of nouns they are and the reason why they occur without an article).

6.3 Descriptive analysis of the bare NPs dataset

The analysis after the retrieval of the parsing patterns focuses on the English part of the data. In addition to the filtering according to the parsing patterns described above, the data was automatically annotated for the part-of-speech tag, number of the noun (i.e. singular vs. plural),[73] and manually checked afterwards. In a second step, the

[73] Due to the nature of the automatic annotation, the distinction between singular and plural nouns is based on morphological criteria (i.e. presence/absence of plural marking). Therefore, singular nouns also include non-count nouns; in terms of terminology, the classification would more precisely be plural vs. non-plural cases.

annotation was refined with more factors that were added manually. These concerned the noun type, countability, reference, and article variability.

The annotation for 'noun type' distinguished between different noun categories that were found in the data, namely abstract nouns (e.g. *safety, peace*), concrete nouns (e.g. *bridges, drift nets*), collective nouns (e.g. *staff, Parliament*), mass nouns (e.g. *money, tuna*), proper nouns (e.g. *Switzerland, Agenda 2000*), abbreviations (both acronyms, e.g. *NATO, CEDEFOP* and initialisms, e.g. *WTO, GNP*), nouns referring to institutions (e.g. *Tübingen University, governments*)[74], and nouns referring to people, i.e. individuals (e.g. *farmers, colleagues, victims, reporters*). Particular attention was paid to the distinction between countable and uncountable nouns, i.e. "entities that can be counted" and "entities that cannot be counted" (Huddleston and Pullum 2002: 334). Examples are given in (10) and (11), where *plate* is considered a count noun and *crockery* a non-count noun. While it is possible to combine numerals with count noun (e.g. *one plate, two plates, three plates*, etc.), this is typically not the case with non-count nouns (e.g. **one crockery, *two crockeries, *three crockeries*). However, many nouns can be considered either countable or uncountable, depending on the noun interpretation, as shown in (12) and (13), respectively. In the former case, *chocolate* indicates an individual unit (i.e. countable noun), while in the latter, it refers to a food substance (i.e. uncountable noun). If there is one lexical item with two different meanings, as in this example, Huddleston and Pullum (2002: 334) talk about *polysemy*, because *chocolate* "has more than one sense".

(10)	We need another plate.	[countable]
(11)	We need some more crockery.	[uncountable]
(12)	Would you like [another chocolate]?	[countable]
(13)	Would you like [some more chocolate]?	[uncountable]

The factor 'reference' has two values, namely *specific* and *generic*. Finally, article variability concerns those cases in which an article could have been used (or not) in the determiner slot.

[74] Depending on the context, the same noun can belong to two different categories. Compare, for instance, the following sentences: *I ask Parliament to support my proposals […]*. (*CoStEP* 2000-03-14.xml) and *The institution is Parliament itself.* (*CoStEP* 1999-03-11.xml), in the former, the noun *Parliament* is a collective noun because the verb *to ask* implies a request, which is normally addressed to an individual or a group of individuals, while in the latter it refers to the parliamentary organization, i.e. the institution.

As described in section 2.4, article use in English differs based on the type of noun that follows (see e.g. Quirk et al. 1985). The distribution of noun types occurring in the bare NPs dataset is given in Figure 6.5.

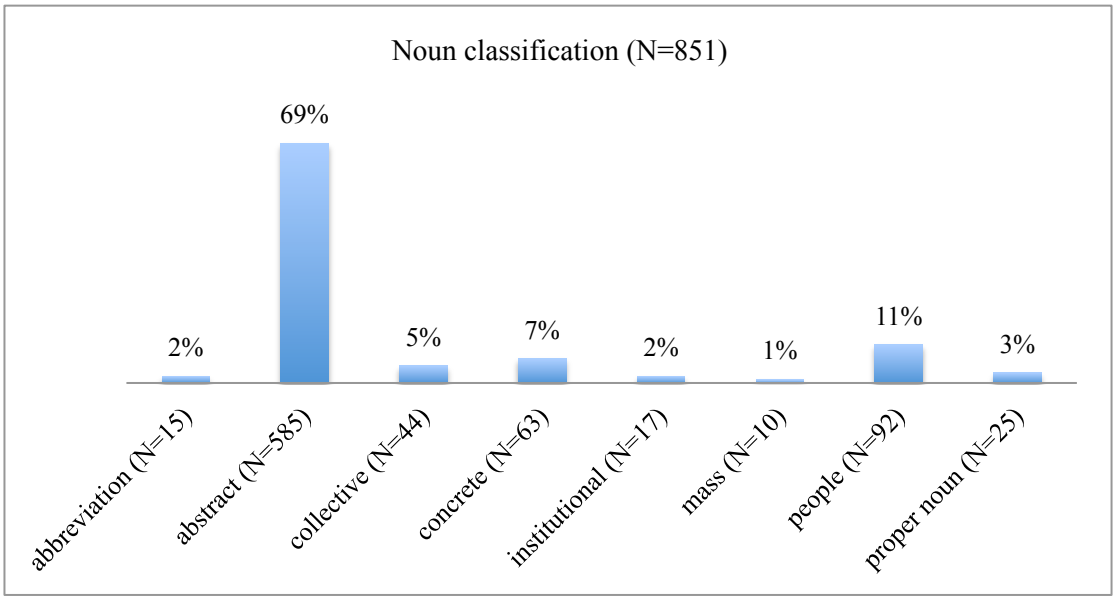

Figure 6.5: Distribution of noun types in the whole dataset of bare NPs.

The graph shows that abstract nouns constitute two-thirds of the data (69%), 11% of the cases contain a noun that refers to people, followed by concrete nouns (7%), collective nouns (5%), proper nouns (3%), abbreviations and institutional nouns (both at 2%), and finally, mass nouns (1%). The high number of abstract nouns is explained by the methodology adopted for the data retrieval, i.e. the use of German as a starting point. As seen in section 2.6, with respect to article use, one of the differences between English and German regards this noun category. In German, abstract nouns are normally preceded by an article, whereas they tend to occur as bare NPs in English (Rowlinson 1994: 87).

Number	
singular nouns	63% (*N* = 537)
plural nouns	37% (*N* = 314)

Table 6.1: Distribution of singular and plural nouns in the bare NPs dataset.

The overview in Table 6.1 refers to the noun number and shows that 63% of the nouns included in the whole dataset of bare NPs are singular, 37% are plural. This is possibly due to the high share of singular cases in the abstract noun category (i.e. 76%).

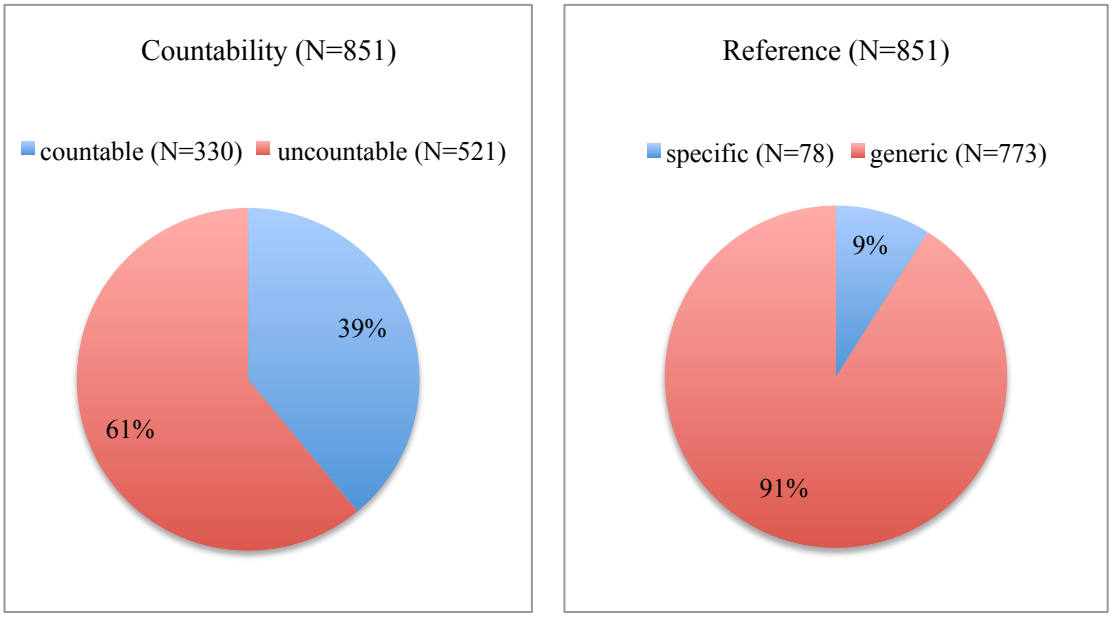

Figure 6.6: Distribution of countability and reference factors in the bare NPs dataset.

Figure 6.6 shows the distribution of countable and uncountable nouns in the pie chart on the left and the distribution of specific and generic cases in the pie chart on the right. The results show that 61% of the whole bare NPs dataset consists of non-count nouns. This finding therefore aligns with previous literature, i.e. that bare NPs mainly occur with uncountable nouns (Quirk et al. 1985: 282). A closer look at the data reveals that the remaining 39% mainly refers to plural nouns that are naturally countable. Furthermore, the results show an almost categorical trend towards the generic reference (i.e. 91%). The findings then support what standard grammars state, namely that English articles are usually omitted in generic noun phrases because these refer "to a whole class rather than to an individual person or thing" (Biber et al. 1999: 265). However, what is striking in this chart is that 9% of the bare NPs have specific reference. It is therefore interesting to see whether the specific cases fall more frequently into the singular or plural data. The distribution is illustrated in Figure 6.7.

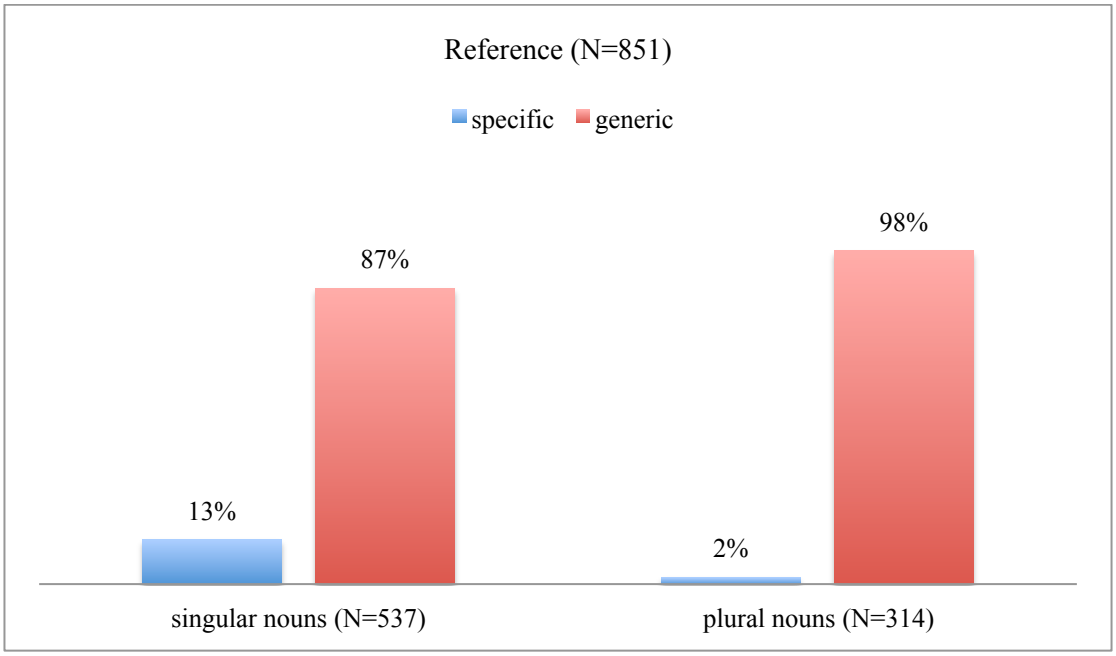

Figure 6.7: Comparison of countability and reference between singular plural nouns.

The results show that the generic reference is more frequent in the plural part of the data (98%) than in the singular part (87%). What is more interesting about the results is that the dataset with singular nouns contains more occurrences with specific reference (13%) than the one with plural nouns (2%). Singular nouns with specific reference refer to proper nouns (e.g. *Parliament, Iran, Switzerland*[75]), which are normally used without an article (Quirk et al. 1985: 288), and abbreviations such as *REACH, ECOFIN, NATO*, which are acronyms. Since these behave similarly to proper nouns, the article tends to be omitted (see e.g. Harley 2004, Callegaro et al. 2019).

In the last part of the analysis,[76] attention was paid to article variability, i.e. those cases in which an article can be added without rendering the sentence implausible or ungrammatical. The results are reported in Table 6.2 and show that the distribution of variable and non-variable cases over the dataset of bare NPs is almost equal with slightly more variable instances. The high number of variability could be

[75] As mentioned in section 2.6, some country names in German require the definite article. Interestingly, *die Schweiz* and *der Iran* are two of these.
[76] Note that throughout this section the sentences with NPs occurring with an article were retrieved from the sub-corpus of English native speakers (i.e. British or Irish), which was created with *Sketch Engine*. [https://www.sketchengine.co.uk]

due to the nature of the selected data, i.e. English bare noun phrases retrieved from German alignments with an article.

Article variability	
variable	54% (*N* = 461)
non-variable	46% (*N* = 390)

Table 6.2: Comparison of variable and non-variable cases in the bare NPs dataset.

A closer look at the distribution of article variability among singular and plural nouns shows that plural nouns allow for more article variability than singular nouns: in three-fourths of the plural data (75%), it is possible to insert an article, while in 25% of cases this is not possible.[77] The results are presented in Figure 6.8.

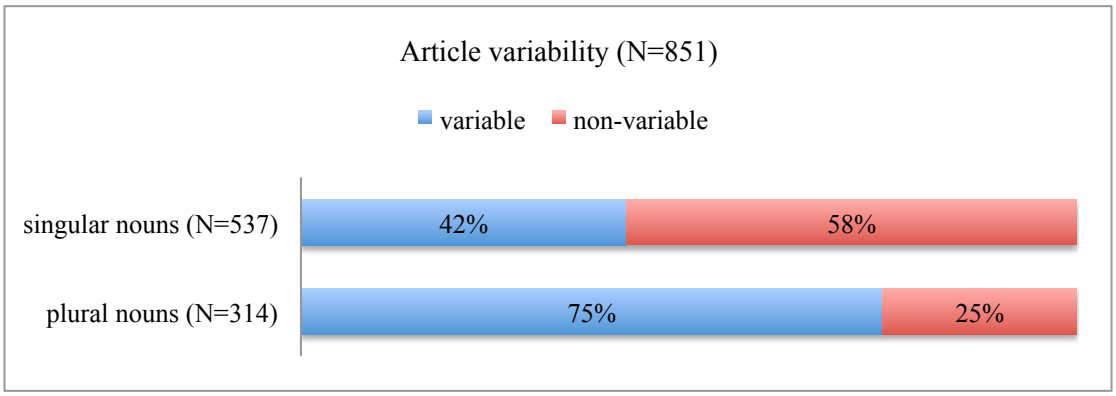

Figure 6.8: Comparison of article variability among singular and plural nouns.

The plural cases which allow for article use are those cases in which NPs change from a generic reference (i.e. article omission) to a specific reference (i.e. article use). Examples of such cases are shown in (14) and (15).

[77] The distribution of article variability among noun type, countability and reference are not presented as for the former there is too much sparse data for all types except for abstract nouns and the latter two are actually influenced by article use and not independent thereof, as section 6.4 will show.

(14)

(a) Ø Rapporteurs should be disinterested and Parliament should adopt this principle in future. (*CoStEP* 1997-04-23.xml)
(b) As you know, the rapporteurs commenced their work long before the horrors of that Dutroux case became known. (CoStEP 1996-12-11.xml)

(15)

(a) I wholeheartedly support this report by Mr. Dover which aims to improve Ø working conditions for young people in Britain. (*CoStEP* 2001-02-13.xml)
(b) […] if this 'Star' Alliance starts to diminish the working conditions and contracts within that […]. (*CoStEP* 1998-03-31.xml)

Sentences (a) occur without an article. However, an article could be used without making the sentence implausible or ungrammatical – what would change is the meaning. In other words, when the definite article precedes the NP, as shown in examples under (b), the context changes from generic to specific. In (14), for instance, the noun *rapporteurs* is non-specific when used as a bare NP, while it acquires specificity when used with the definite article. Thus, the use or omission of an article with plural nouns activates or deactivates the specificity/genericness feature. On the other hand, the sentences that were coded as non-variable are those where the given context is enough to understand that the reference can only be generic, as in (16) – (18), or when the item in question is part of an institutional noun, as in (19),[78] or event nouns, as in (20),[79] and is therefore a fixed expression and does not require an article.

(16) To improve Ø things we need to change the quality of government. (*CoStEP* 1992-12-16yml)
(17) This summit was a start but much has to be done to turn Ø words into reality, and we, Members of this House, must be part of that new world. (*CoStEP* 2000-04-11.xml)
(18) I am strongly in favour of consumers being able to obtain Ø secure, safe, reliable and sustainable supplies of gas and electricity at reasonable prices. (*CoStEP* 2008-06-19.xml)
(19) […] because of the work of national and international bikers' rights organisations coordinated by the Federation of Ø European Motorcyclists, or FEM. (*CoStEP* 1996-06-18.xml)
(20) […] the report proposes that within Parliament we have a joint meeting of the Conference of Ø Committee Chairs and […]. (*CoStEP* 2005-11-30.xml)

[78] Other examples are *the Court of Ø Auditors, the Council of Ø Ministers, the House of Ø Commons, the Committee of Ø Independent Experts, the Committee on Ø Petitions, the European Year of Ø Disabled People*.
[79] Another example is *the Conference of Ø Presidents*.

Other cases that are not likely to occur with an article are presented in (21) and (22). Since *amendments* and *articles* are postmodified by numerals and are written with capital letters in the transcripts, they are considered specific because they refer to specific amendments and specific articles. Thus, the speakers consider these nouns proper nouns and, therefore, they do not require an article.[80]

(21) I supported Ø Amendments Nos 14 and 179 of the above report. (*CoStEP* 2001-05-31.xml)
(22) However, all proposals referring to agricultural policy must respect Ø Articles 32 to 38 of the Treaty and […]. (*CoStEP* 2002-05-30.xml)

Singular nouns exhibit lower variability (42%). Closer inspection of the data reveals that the cases in which an added article would be considered ungrammatical mainly refer to non-count nouns with generic reference. Examples are provided in (23) – (26).

(23) […] and to ensure that, above all else, Ø jingoism is never, ever, heard again in the European Union. (*CoStEP* 1996-06-06.xml)
(24) These include definitions of what constitutes a measure to protect Ø public health and is […]. (*CoStEP* 2003-02-10.xml)
(25) Whilst it is true that Ø terrorism may have an environmental impact, for examples the release of pathogens or […]. (*CoStEP* 2002-04-08.xml)
(26) Society recognizes Ø marriage not on moral grounds but because it recognizes on a rational basis the stability it affords society socially and […]. (*CoStEP* 2000-03-14.xml)

Cases with specific reference are all proper nouns, such as country names, as in (27), institutional nouns, as in (28), acronyms, as in (29), and nouns that are considered proper nouns, as *Parliament* in (30) and *Agenda 2000* in (31).

(27) Madam President, on 9 July, Ø South Sudan will declare independence as an English-speaking new African state. (*CoStEP* 2011-06-07.xml)
(28) Only this weekend there are reports from scientists in such places as the Dutch Institute of Science and Health and Ø Tübingen University, which […]. (*CoStEP* 1996-06-18.xml)
(29) Ø UKIP supports the repeal of all EU legislation. (*CoStEP* 2007-09-05.xml)

[80] A test search in *Sketch Engine* showed that there is a strong tendency for these NPs to be used as bare nouns (out of 178 cases of *Amendments Nos…* no cases were used with an article, while there was only one case of *articles* post-modified by numerals against a set of 59 instances).

(30) Following on from what he has said, Ø Parliament has been told by the Commission that it cannot undertake any action […]. (*CoStEP* 1997-07-16.xml)
(31) I wish Mr. Flynn all success in pursuing Ø Agenda 2000, which is the only way […]. (*CoStEP* 1998-04-01.xml)

Finally, four nouns do not require an article because they are included in fixed expressions, i.e. *part*, as in (32), *favour*, as in (33), *fact*, as in (34), and *place*, as in (35).

(32) National ministers, who are Ø part of national governments, accountable to national parliaments. (*CoStEP* 1997-05-12.xml)
(33) I hope you will listen to the swelling chorus in this House in Ø favour of reform. (*CoStEP* 2007-01-16.xml)
(34) The terrible risk is that, despite the talk about building Europe, it may in Ø fact be undermined […]. (*CoStEP* 2010-12-15.xml)
(35) I would add […] that the meeting that is taking Ø place today between Mr. Giscard d'Estaing and the heads of the political groups should not […]. (*CoStEP* 2002-02-06.xml)

On the question of article variability with singular nouns, the evidence reveals that articles are variable (i.e. an article could be added without making the sentence implausible or ungrammatical) in two contexts. Firstly, as attested by Harley (2004) and Callegaro et al. (2019), article use with initialisms tends to be variable. Examples are given in (36) and (37).

(36) So I am not throwing it out of the window, I am asking us to review it in Ø WTO. (*CoStEP* 2001-10-04.xml)
(37) This report, however, goes well beyond the immediate ambitions of even those who drive Ø ESDP at the moment. (*CoStEP* 2006-11-15.xml)

Secondly, articles could be used in those contexts in which the item in question can semantically be both a countable and uncountable noun. However, the presence or omission of the article is relevant for the meaning the noun conveys. Nouns that are included in this category are, for instance, *legislation, taxation, opportunity, negotiation, responsibility, reform, agreement, clarification, effort, confidence,* and *production*. Examples are the following:

(38)

(a) The influx of people seeking asylum […] calls into question the situation in certain of these countries and makes it urgently necessary for them to introduce Ø effective legislation to protect their ethnic minorities. (*CoStEP* 1999-01-14.xml)
(b) I really hope that when the legislation comes into force, in 1998, the car manufactures will ensure that the vehicles […]. (*CoStEP* 1996-09-18.xml)
(c) The European Parliament must and should use these new powers to push forward an even stronger legislation in the area of […]. (*CoStEP* 1999-11-15.xml)

(39)

(a) The formulation of the common position would have simply allowed the employers to walk away from attempts to reach Ø negotiation on annualisation of working time and […]. (*CoStEP* 1999-11-03.xml)
(b) In all honour to everybody concerned in the negotiation, the Council and all Members of the delegation, they have made […]. (*CoStEP* 2000-05-16.xml)
(c) There comes a time within a negotiation when the kid gloves must come off, when action must be taken. (*CoStEP* 2002-04-10.xml)

(40)

(a) This is a spy base set up by Ø agreement between the British Government and the USA in 1948 among other purposes to monitor […]. (*CoStEP* 1998-09-16.xml)
(b) I shall not oppose the agreement but there are two political points that need to be made. (*CoStEP* 1996-04-16.xml)
(c) […] it is disgraceful that the Council of Ministers is unable to reach an agreement on a compensation package. (*CoStEP* 1996-06-06.xml)

The nouns in question denote abstract concepts and are uncountable in their primary sense, as shown in examples (a). However, they can also denote a secondary countable sense. In (38), the noun *legislation* can indicate either an act of making or creating laws or a law enacted by a legislative body. In (38)a, the noun is considered uncountable, while in (38)b and (38)c it is considered countable, because a legislative body might create many laws (i.e. legislations). Likewise, the word *negotiation*, in (39), can denote either an act or process of negotiation or the result of negotiation. The former relates to an uncountable meaning, as in (39)a, the latter a countable one, as illustrated in (39)b and (39)c. Finally, in (40), the noun *agreement* can refer to either the act of agreeing or an arrangement accepted by two or more parties (or a contract or document with details of an agreement). Similar to the previous nouns, in (40)a the noun is uncountable, while in (40)b and (40)c it is countable. When discussing count and non-count nouns, Huddleston and Pullum (2002: 337) pay attention to the distinction between abstract non-count nouns and event or result count

nouns (i.e. *Permission is required* vs. *Two separate permissions are required* and *Necessity is the mother of invention* vs. *Edison was honoured for three separate inventions*). The examples provided by Huddleston and Pullum (2002) refer to count nouns only in their plural form and do not mention article use. However, corpus evidence shows that an article with the singular can be used to make the same semantic distinction. In other words, since these nouns originally occur as bare NPs, the nouns refer to the non-count sense; as marked in the examples above, a definite or an indefinite article could be used, but this might entail a change of meaning (i.e. from abstract non-count noun to abstract-result count noun).

In terms of article variability with abstract nouns, a closer look at the data reveals that there is one construction in particular in which an article is expected but does not always occur. The construction in question consists of an abstract noun followed by the preposition *of*, hereafter called the *of*–CONSTRUCTION. Quirk et al. (1985: 286-287) claim that abstract nouns are normally realised as bare NPs when they occur without modification but usually require an article when postmodified by an *of*-phrase construction, as shown in (41):

(41)
(a) She's studying European history.
(b) She's studying the history of Europe.
(c) She's studying *history of Europe.

They state that the definite article is used in (41)b "because the effect of the *of*-phrase is to single out a particular subclass of the phenomenon denoted by the noun" (Quirk et al. 1985: 286). Thus, the *of*–CONSTRUCTION changes the meaning from generic to specific. Furthermore, according to them, it is not possible to have an NP postmodified by an *of*-phrase without an article, as shown in (41)c. However, they only present three instances using the same abstract noun, i.e. *history*. In a collostructional analysis (Callegaro and Clematide 2017), the non-occurrence of the expression *Ø history of* in *CoStEP* was confirmed, but a quick look at other corpora reveals that it is definitely possible. For instance, the NOW corpus contains the example *I hope not, but if Ø history of Europe is anything to go by reality looks worse.* Furthermore, other instances are found in COCA, for example: *My own view, as somebody who has studied Ø history of international criminal law […]* and *The interview protocol covered Ø history of substance use, abstinence, help-seeking*

attempts […]. It is therefore probable that *CoStEP* does not include these instances due to a subject limitation: it is a fairly specialised corpus (i.e. parliamentary discourse). However, a search over our corpus shows that cases postmodified by the preposition *of* and occurring without an article appear with different lexemes. Moreover, it shows that the definite article in this construction is used variably, as shown in the following examples.

(42)
- (a) […] and bring forward measures to prevent Ø loss of services to those regions concerned and prevent job losses. (*CoStEP* 1998-07-01.xml)
- (b) The loss of one container of cigarettes, it is reported, costs us £800,000 in lost revenue. (*CoStEP* 1997-03-13.xml)

(43)
- (a) However, it also brings certain obligations - particularly the need to guarantee Ø universality of services. (*CoStEP* 2000-12-13.xml)
- (b) […] this House itself reiterated the principle of the universality of human rights and non-discrimination as a basis […]. (*CoStEP* 2010-12-16.xml)

(44)
- (a) Mr. Carreia de Campos is concerned about Ø movement of workers. (*CoStEP* 2011-04-06.xml)
- (b) While the American government has halted the movement of potentially dangerous nuclear materials, BNFL is […]. (*CoStEP* 2001-10-22.xml)

(45)
- (a) It is also necessary to ensure that […] and that there is Ø real equality of opportunity built in. (*CoStEP* 1998-04-29.xml)
- (b) […] I will promote the equality of gender in various ways, but most visibly and at the outset by it being apparent in […]. (*CoStEP* 2002-01-15.xml)

These instances support the assumption that the meaning depends on the absence or presence of an article – i.e. generic or specific reference – as shown in examples (a) and (b), respectively.

As the final point of the descriptive analysis of bare NPs, article variability among the parsing patterns used in the data retrieval is presented in Figure 6.9. Overall, there is no striking difference and the distribution is fairly homogenous. The most deviant pattern – i.e. PASSIVE SUBJECT PHRASE – could be explained based on the lower underlying counts, although showing significant difference in a chi-square

test.[81] While interesting to investigate in further research, parsing patterns will not be analysed in more detail and, as previously mentioned, the focus will be on the internal properties of an NP.

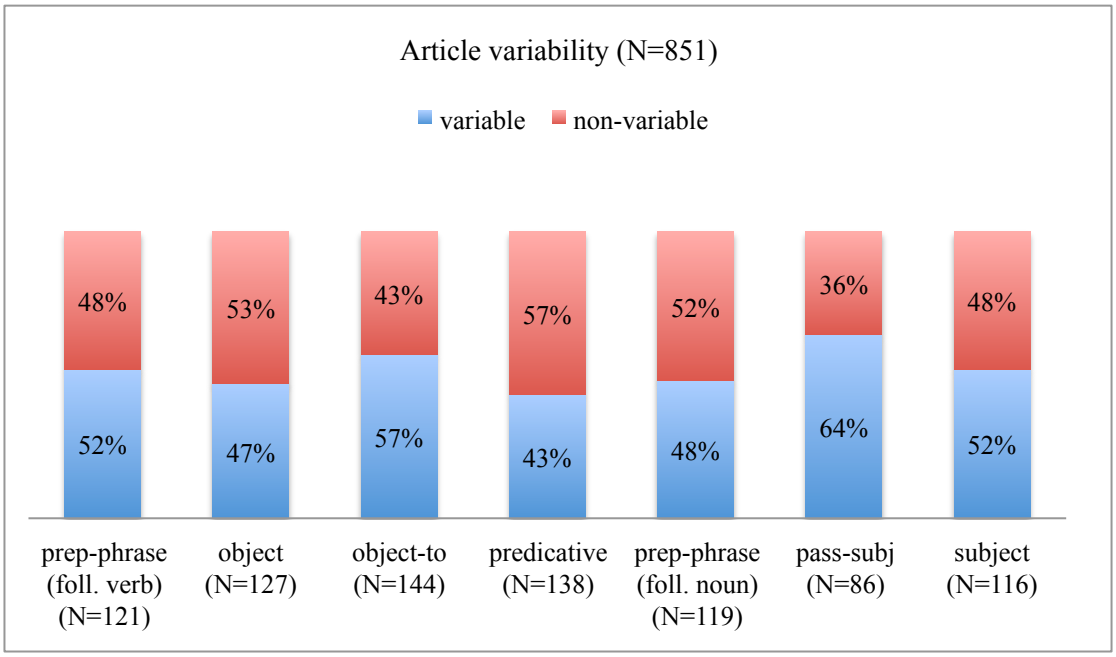

Figure 6.9: Distribution of variable and non-variable cases among the retrieval patterns.

Taken together, these results provide important insights into the nature of bare NPs. The most relevant observation is the fact that corpus evidence confirms what Quirk et al. (1985: 274) and Biber et al. (1999: 261) describe in their grammars, namely that articles are not normally used with uncountable nouns and with plural countable nouns. Moreover, for the singular nouns, the results corroborate the tendency for article omission to mainly occur with non-count nouns with generic reference (Quirk et al. 1985: 282). As previously mentioned, it is important to remember that these instances were retrieved using German as the starting point. Contrary to English, German normally requires an article in front of abstract nouns (Rowlinson 1994: 87). This might, in turn, explain the high number of abstract nouns in the dataset. One particular case discovered via the data-driven analysis is the structure in which a bare abstract noun is postmodified by an *of*-phrase. Contrary to what previous literature claims possible, variability is found in the corpus. Therefore, this specific

[81] Besides the NON-FINITE PHRASE, the difference between the PASSIVE SUBJECT PHRASE and the other patterns is significant at $p<0.05$.

construction is worth investigating in more detail, in comparison to the construction in which an abstract noun occurs without postmodification. The analysis will be discussed in the following section.

6.4 Abstract nouns vs. *of*–CONSTRUCTION

The case study for this analysis consists of two subsets of newly retrieved data: one includes abstract nouns postmodified by an *of*-phrase (i.e. *of*–CONSTRUCTION), the other one contains cases of abstract nouns that are not postmodified. Importantly, the new data sample includes both bare NPs and noun phrases preceded by an article. Besides a detailed analysis of the *of*–CONSTRUCTION with abstract nouns, the data therefore allow to provide a more in-depth evaluation of variable article use with abstract nouns, based on empirical evidence, that will be the basis to build a Construction Grammar model.

For this follow up-analysis, the singular abstract nouns (a set of 284) from the dataset of bare NPs were used as the starting point for the analysis.[82] A frequency list of these nouns was necessary to select the lexical items to be included in the close-up. For this frequency list, the filters described in section 6.2 were applied and only the nouns that occurred at least 50 times in the corpus were taken into consideration.[83] The total number of nouns (types) included in the analysis was 196. During the retrieval process, since a total random retrieval would have had a strong bias towards the most frequent nouns, a large lexeme variety was prioritized instead and 6 instances per lemma were retrieved (a total of 1.176 sentences).[84] Despite the filters, the retrieval again included false positives, namely other elements occupying the determiner slot (e.g. possessives, demonstratives). Additionally, alignment errors resulted in the retrieval. In other cases, the automatic annotation coded an article when this was related to a fixed expression (e.g. *a little more evidence*). Further fixed expressions were also excluded from the analysis (e.g. *in control, in effect, to date, to lose sight, to make use*). Finally, genitives and incomplete sentences were deleted from the dataset as well. The final total number of instances included in the dataset was 976. As mentioned before, the sample was first divided into two groups: the

[82] The complete list of nouns included in the dataset of bare NPs is presented in Appendix B.
[83] The complete frequency list of the nouns is presented in Appendix C.
[84] The first two nouns on the frequency list (i.e. *policy* and *industry*) occur twice as often than the next 20 nouns.

former without an *of*-phrase as postmodifier (the large majority of cases), the latter with an *of*-phrase (i.e. *of*–CONSTRUCTION). In order to have a better data classification, the dataset was further subdivided according to the presence of the *of*-phrase as well as the article. Table 6.3 provides a summary of the four constructions.[85]

Construction		**Example**
C1	[art] [abstract N] [Ø]	[the] [industry] [Ø]
C2	[Ø] [abstract N] [Ø]	[Ø] [industry] [Ø]
C3	[art] [abstract N] [of]	[the] [use] [of]
C4	[Ø] [abstract N] [of]	[Ø] [use] [of]

Table 6.3: Constructions investigated in the follow-up analysis.

The distribution of the constructions in the whole dataset is shown in Figure 6.10.[86] The results show that the construction in which an abstract noun occurs as bare NP, i.e. C2, is the one that occurs more frequently (62%); this is followed by C1, the construction with an abstract noun following an article (19%), and by the constructions belonging to the *of*–CONSTRUCTION: C3, the one occurring with an article (16%), and C4, the one in which an article is omitted (3%). Therefore, the findings reveal that abstract nouns without a postmodifying *of*-phrase are more likely to occur in a bare NP.

[85] Note that these constructions do not show the possibility for the noun in question to have an adjective as premodifier.
[86] Due to the low frequency of cases preceded by the indefinite article, this variant was combined with the definite article.

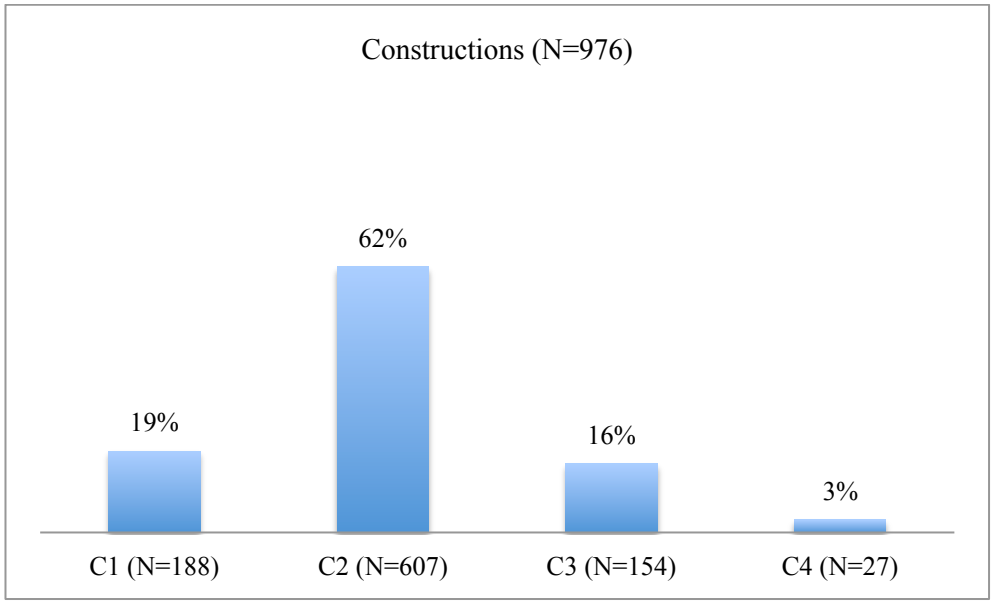

Figure 6.10: Frequency of constructions in the whole dataset.

A chi-square test was used to test whether the difference between the subset without an *of*-phrase (i.e. C1 and C2) and the one with an *of*-phrase (i.e. C3 and C4) was significant with regard to article use. The difference was significant (at $p<0.05$). Corpus evidence thus supports Quirk et al.'s (1985) observation, namely that abstract nouns are normally realised as bare NPs but take an article when postmodified by a prepositional phrase (in this case with *of*).[87] However, Figure 6.10 also shows that there is more variation than what Quirk et al. (1985) claim. In fact, contrary to what they state, 3% of the whole dataset appear with an *of*-phrase as postmodifier and without an article.

The current data sample was annotated for article use (presence vs. absence), countability (countable vs. uncountable), reference (specific vs. generic) and premodification. In the data, different types of premodifying elements occur, i.e. a noun (e.g. *the codetermination policy*), a proper noun (e.g. *the Kosovo crisis*), an adjective (e.g. *scientific development*), an acronym (e.g. *EU law*), a numeric element (e.g. *the 1996 discharge*), or a combination of them (e.g. *EU development policy*). However, due to the low frequency of some premodifiers, they were all combined

[87] Two other constructions were found with different prepositions, namely *for* and *in* (e.g. *Ø support for Islamic fundamentalist groups* and *Ø oil exploration in the Ecuadorean rain forest*). These were not taken into account for further analysis because they might behave similarly.

together. Figure 6.11 shows the distribution of premodifying elements among the four constructions.

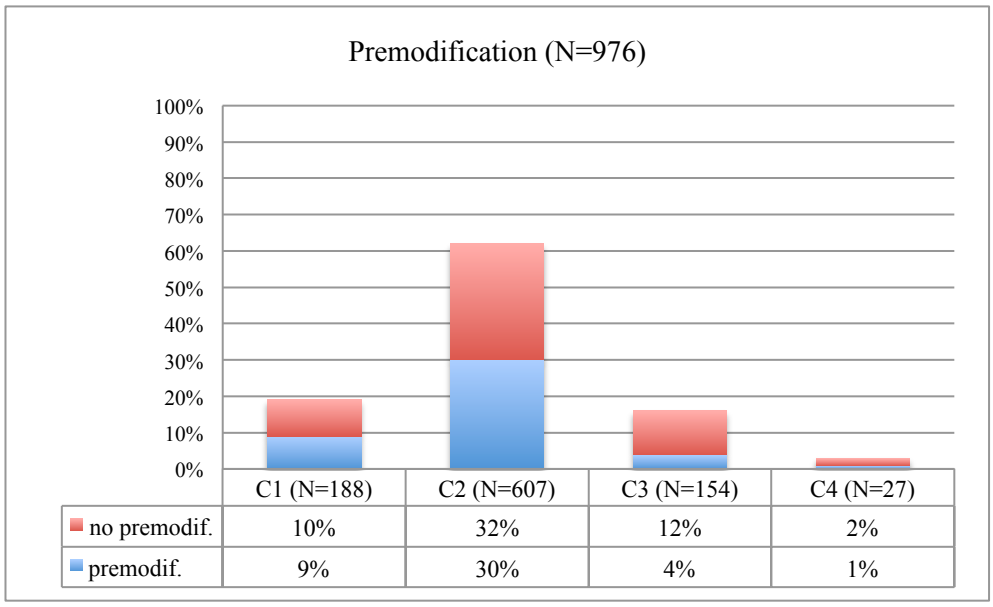

Figure 6.11: Distribution of premodifying elements among constructions.

The graph shows that for the non-postmodified constructions (i.e. C1 and C2) premodification is equally distributed, while the cases with an *of*-phrase (i.e. C3 and C4) occur more often without premodification. The four examples below show a premodified noun in each construction.

(46) We support the join resolution, but, at the same time […]. (*CoStEP* 2003-09-04.xml)
(47) It promotes Ø appropriate future legislation to give more flexibility in the implementation of TRIPS to encourage […]. (*CoStEP* 2001-10-04.xml)
(48) The forthcoming ratification of the European Constitution, despite its rejection in two referendums, is undemocratic […]. (*CoStEP* 2008-01-17.xml)
(49) I welcome Ø improved coordination of economic and fiscal policy in Europe, but strongly oppose the short-term vision under […]. (*CoStEP* 2011-06-23.xml)

A chi-square test was used to test whether the difference in article use across premodified and non-premodified cases was significant or not. The results are reported in Table 6.4, which distinguishes between the group of instances without an *of*-phrase (i.e. C1 and C2) and with an *of*-phrase (i.e. C3 and C4). For both subsets, the

statistical test shows that premodification is not significant (i.e. p>0.05). Thus, article use is not influenced by the presence or the absence of a premodifier.

	χ^2	*p*
premodif. vs. no modif. C1 – C2	0.10	0.75
premodif. vs. no modif. C3 – C4	0.01	0.92

Table 6.4: Significance of premodification as a factor predicting article use.

The next investigated feature was the countability of the nouns. Figure 6.12 gives the distribution of count and non-count nouns among the four constructions. What stands out in the graph is that there is a clear tendency among the nouns occurring without an *of*-phrase as postmodifier.

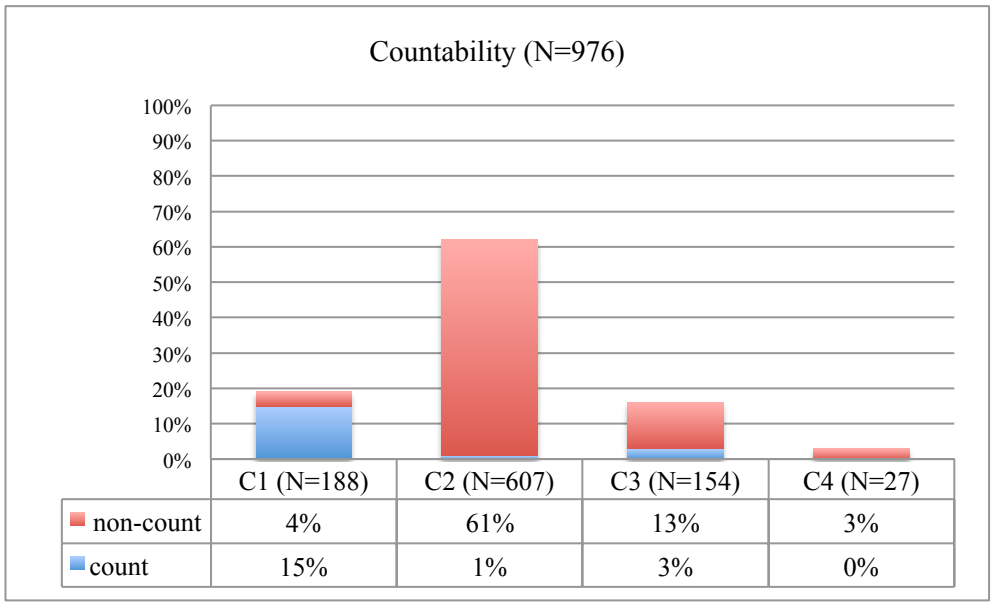

Figure 6.12: Distribution of count and non-count nouns among constructions.

The results confirm that bare abstract nouns are mainly uncountable (i.e. C2), while countable abstract nouns show a strong preference towards article use (i.e. C1). However, from the graph, one can see that this distribution is not categorical: 4% in the C1 subset includes non-count nouns preceded by an article. A closer look at the data reveals that part of these nouns are exclusively uncountable, as shown in (50) and

(51), while the rest concerns abstract nouns that can naturally occur with an article as well, as shown in (52), (53), and (54).

(50) […] and the work they had been engaged in had been of a European nature. (*CoStEP* 1993-03-22.xml)
(51) She gives the bad news in the nicest possible way. (*CoStEP* 1999-01-12.xml)
(52) Above all, they are entitled to have their data in the electronic communications space protected. (*CoStEP* 2009-05-05.xml)
(53) […] those who are legally within our borders will be treated with the same dignity and respect that we would like […]. (*CoStEP* 2010-12-13.xml)
(54) What we need are the measures outlined in our resolution in order to see the real growth and the real jobs that we need […]. (*CoStEP* 1996-06-18.xml)

The C2 subset presents a small amount of cases (i.e. 1%), in which abstract nouns that are countable are used in bare NPs. However, these cases cannot be considered genuine bare NPs, because there are other factors outside the NP influencing article use. As shown in (55) and (56), the noun in question is a complement of an *of*-phrase. In examples (57) and (58) the noun phrase behaves as an idiomatic expression that involves the preposition *at*, the absence of an article, a (locative) premodifier, and the lexeme *level*. In *CoStEP*, more examples are found, such as *at Ø national level*, *at Ø world level*, *at Ø Community level*, *at Ø veterinary level*.

(55) If you are suggesting we give up common law and habeas corpus for that sort of Ø European system, my answer to you is no […]. (*CoStEP* 2007-05-22.xml)
(56) […] which is what I am used to in my type of Ø area – the Highlands and Islands of Scotland. (*CoStEP* 1997-09-18.xml)
(57) That is something that we should be considering seriously at Ø European level. (*CoStEP* 1999-03-09.xml)
(58) […] in practical terms it is still very rare to find true effective competition at Ø local level for the domestic consumer. (*CoStEP* 1999-01-27.xml)

Turning now to C3 and C4, i.e. the constructions occurring with an *of*–CONSTRUCTION, it is possible to note that C3 occurs more often with non-count nouns, while C4 only occurs with uncountable nouns; examples are given in (59)[88] and (60), respectively. However, C3 shows some variability. (61) and (62) are two examples of countable nouns.

[88] This instance might be ambiguous. However, based on the context in which the sentence was uttered, *reform* refers to the process of creating a reform and was therefore annotated as uncountable.

(59) […] that a statute for Members should form part of the successful reform of the European Union institutions. (*CoStEP* 2003-01-14.xml)
(60) It is fitting that Ø freedom of information legislation should be adopted under the presidency of a Member State […]. (*CoStEP* 2001-05-02.xml)
(61) […] to establish whether the policy of restricted access has assisted in meeting the objectives of the […]. (*CoStEP* 2003-06-03.xml)
(62) On the issue of the development of industrial sites, the Commission should not wring its hands but say […]. (*CoStEP* 2001-01-16.xml)

Table 6.5 reports the results of the chi-square test that was applied to see whether the difference between the two groups is statistically significant.

	χ^2	*p*
count vs. non-count C1 – C2	565.1	0.0001***
count vs. non-count C3 – C4	5.6	0.018***

Table 6.5: Significance of countability vs. uncountability.

The results are not surprising; in fact, they show that the difference is significant in both groups (at $p<0.05$).

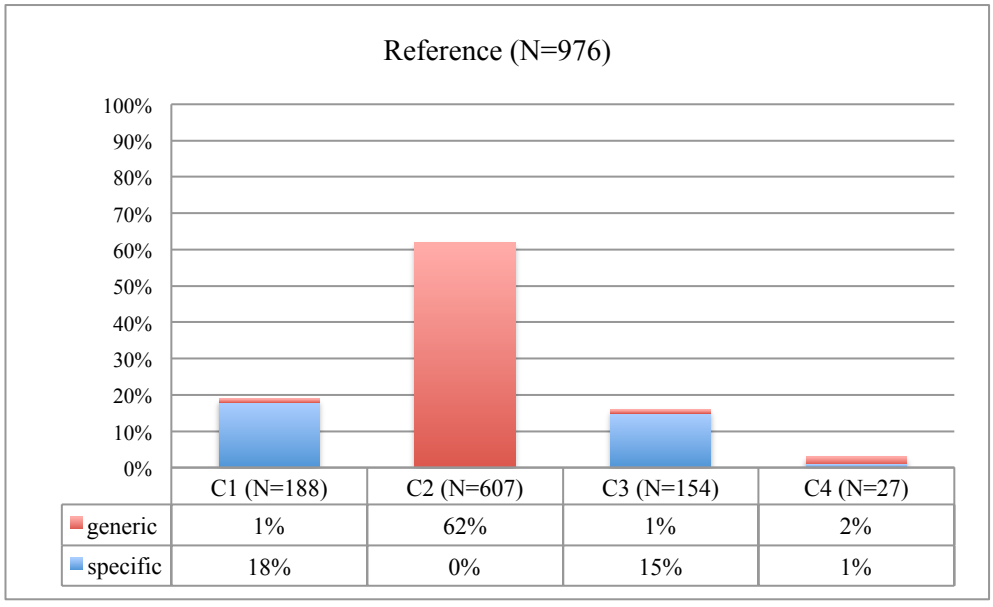

Figure 6.13: Distribution of specificity and genericness among constructions.

In the following paragraphs, the final factor on article use is discussed, i.e. the specific and generic reference of an NP. The distribution of specific and generic cases

among the constructions is given in Figure 6.13. It is apparent from the figure that article use and article omission follow clear tendencies: namely, the cases occurring with an article are almost always specific, while bare cases are mostly used in a generic sense. This is valid for C1, C2, and C3, as shown in (63), (64), and (65), respectively.

(63) I agree with one of the previous speakers who said that the consumer must have the choice and must be able to […]. (*CoStEP* 2006-10-11.xml)
(64) The creation of an internal market in electricity should provide consumers of electricity with Ø real choice. (*CoStEP* 2000-03-29.xml)
(65) We must look towards common solution to Europe's problems, whilst respecting the choice of each EU Member State by […]. (*CoStEP* 2010-10-20.xml)

What is striking about Figure 6.13 is that C4 shows variability. Let us now consider this construction in more detail to see in which contexts this occurs. As already shown in Figure 6.10, C4 does not occur very frequently (i.e. 27 cases). There is one particular instance with specific reference. The sentence is shown in (66)a. The abstract noun *enlargement* is postmodified by the expression *of the European Union* and occurs without an article, which could be added without changing the meaning of the whole construction. The counterpart instance found in the data sample occurring with an article is given in (66)b.

(66)
(a) Your first priority, rightly, was to meet the timetable for Ø enlargement of the European Union. (*CoStEP* 2003-07-01.xml)
(b) […] I am pleased to participate in this important debate on the enlargement of the European Union and in particular […]. (*CoStEP* 2001-09-04.xml)

It could be argued that this might be an occasional occurrence, but a search over the corpus reveals that there are more similar cases, as shown in (67) – (70).

(67) […] there is much in the Brok report that we can support, in particular the importance attached to Ø enlargement of the European Union. (*CoStEP* 2000-11-29.xml)
(68) In talking about Ø enlargement of the European Union, it is important in providing exchanges […]. (*CoStEP* 1998-01-28.xml)
(69) Ø Enlargement of the European Union would give us a common border with Ukraine. (*CoStEP* 1998-03-11.xml)

(70) The main argument in favour of these amendments and of a report at the end of five years actually concerns Ø enlargement of the European Union […]. (*CoStEP* 2000-12-11.xml)

Thus, we are dealing with a clear case of article variability; that is, the presence or absence of the definite article does not determine the distinction between specific or generic reference. Interestingly, the expression *enlargement of the European Union* is seen as something unique within the European Parliament. In other words, it is likely that speakers have started to use it as one unit and not as a free phrase. This phenomenon is similar to what was discussed in Chapter 5, in which the collective noun *Parliament* behaves as a proper noun, in contrast to *the European Parliament*, whose article use makes the noun behave like a common noun. The difference with the current case is that the NP as a whole is considered as a proper noun (i.e. including all the elements of the *of*–CONSTRUCTION).

The cases included in the C4 data subset express genericness. A few examples are given in (71) – (74).

(71) […] with that of the nuclear industry, in terms of accidents involving Ø loss of life. (*CoStEP* 2011-03-23.xml)
(72) If Ø taxation of energy is to be used for pursuing environmental aids, then the policy has to be transparent and […]. (*CoStEP* 2000-09-20.xml)
(73) If we are not careful, we might end up with Ø increased mobility of exploitation, which would hardly achieve either […]. (*CoStEP* 2003-06-18.xml)
(74) The problem I am seeking to address in this question is that Ø concentration of media ownership sometimes causes problems […]. (*CoStEP* 1999-02-10.xml)

Like C1 and C2, the results therefore show that C3 and C4 differ because the presence and omission of an article determines the specific and generic reference, respectively. Put differently, there are two distinct constructions with two different meanings. It therefore comes as no surprise that the difference proves significant in a chi-square test (at $p<0.05$).

	χ^2	p
specific vs. generic C1 – C2	789.00	0.0001***
specific vs. generic C3 – C4	164.18	0.0001***

Table 6.6: Significance of specificity vs. genericness.

This analysis has shown that abstract nouns follow clear tendencies regarding the influence of countability, reference, and postmodification with an *of*-phrase on article use. Corpus evidence therefore provides a strong base on which Construction Grammar can be applied. The next section moves on to characterize and discuss these tendencies with the help of Construction Grammar and presents a CxG model. In particular, the constructional analysis focuses on the base form of an NP, which only includes the determiner and a noun and excludes other factors, such as syntactic function, context, or discourse. As mentioned in section 3.1, the focus is on the semantic properties that are present in the constituents of an NP and on the semantic properties that are projected onto the construction as a whole. Furthermore, due to non-significant results with regard to premodification and for sake of simplification, this factor is not taken into account. The investigation thus considers abstract nouns without modification and with one particular case of postmodification (i.e. with the preposition *of*). In addition, since plural nouns also showed article variability (see section 6.3), they are also addressed in the following constructional analysis. The entire CxG modelling is strictly based on empirical evidence. Langacker (1991: 6-7) stresses that a bottom-up approach is a natural solution to the problem of specifying which elements are allowed in a particular construction, i.e. that "a high-level schema describing a broad generalization does not exist in isolation; rather it is one node in a network that also includes subschemas corresponding to special cases of the general pattern, which may in turn have subschemas, and so on." On the other hand, "abstraction can be carried to any degree supported by the data" (Langacker 2005: 144).

6.5 A CxG model of variable article use

Within the constructional hierarchy, the CxG diagrams of non-postmodified NPs are presented here are at what Traugott (2008: 236) calls the micro-level, which includes

"individual construction types", while postmodified NPs (i.e. *of*–CONSTRUCTION) are analysed at the meso-level, which refers to "sets of similarly-behaving specific constructions". As previously mentioned, the description of the micro-level and meso-level constructions derives from corpus evidence (i.e. the data presented and analysed in sections 6.3 and 6.4). From a constructional point of view, a data-driven approach allows for a bottom-up method: the constructional analysis is based on the lowest level of abstraction – i.e. the construct level – and moves up to a higher level – i.e. the micro-level or the meso-level. Table 6.7 illustrates the hierarchical organization of non-postmodified NPs. The hierarchical structure begins at the bottom with an actual clause and moves up to more abstract levels, until reaching the highest level of abstraction (i.e. the macro-level), which consists of the basic form of an NP. The *of*–CONSTRUCTION is analysed at the meso-level, whose macro-construction consists of a nominal postmodified by a prepositional phrase. The *of*–CONSTRUCTION is thus a more specific version thereof.

Macro-level	[basic NP]
Meso-level (3)	[[determiner] + [noun]]
Meso-level (2)	[[article] + [noun]]
Meso-level (1)	[[Ø] + [proper noun]], [[the/Ø] + [plural count noun]], […]
Micro-level	[[Ø] + [Switzerland]], [[the/Ø] + [colleagues]], […]
Construct-level	*Although Switzerland is not a Member State, Swiss people are informed Europeans.*

Table 6.7: Representation of the constructional hierarchy taken into account in the current analysis.

The following diagrams are based on the constructional models of the Determination Construction designed by Fillmore (1988) and Fried and Östman (2004b), introduced in section 3.2. In particular, the diagrams use the basic constructional structure (i.e. the representation of the elements' slots within a construction) from the former and the semantic properties of the constituents included in the slots from the latter. Moreover, the constructional representations adopt the same binary opposition features of both Fillmore (1988) and Fried and Östman (2004b). The first attempt of the CxG description regards plural count nouns, as their difference in meaning with respect to article use is more straightforward than in other constructions. The model is

given in Figure 6.14. The presence and absence of the definite article in fact determines the specificity and genericness of a plural count NP.

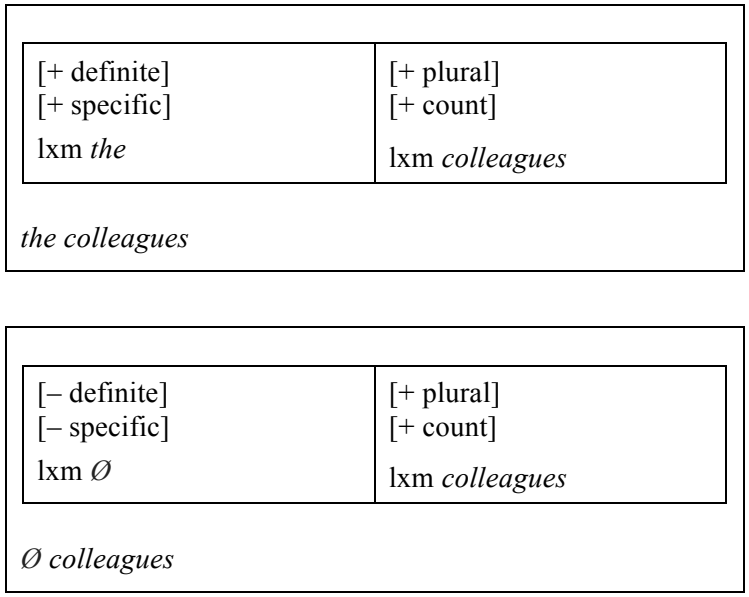

Figure 6.14: Constructions of plural count nouns.

Both diagrams consist of an outer box that represents an NP and contain two smaller boxes (or slots): the box on the left refers to the article and its values (i.e. determiner slot), while the box on the right includes the properties that a noun has when co-occurring with or without an article (i.e. noun slot). The upper construction relates to NPs that take a definite article and then represents a plural count noun whose specificity is marked positively. The construction as a whole inherits the features of the individual parts. The features are, therefore, [+definite] and [+specific] from the determiner, and [+plural] and [+count] from the noun. By contrast, the second construction describes article omission and represents an NP with a plural count noun with negative specificity. Hence, the properties in this case are [–definite], [–specific], [+plural] and [+count].

The difference between these two constructions thus regards both the definiteness and the specificity feature, which are due to article use. In other words, a plural NP can have either specific or generic reference and can be either definite or non-definite, depending on the presence or absence of the article. The tokens used as examples, marked as *lxm* (i.e. *lexeme*), form the expressions *the colleagues* and *Ø colleagues*. Hence, the noun *colleagues* is one element of the plural count noun

construction and can, therefore, have its own properties. However, in order to establish whether the reference is specific or generic, it needs the features from the definite article or from its omission. It is, in fact, not possible to know whether a plural count noun alone is generic or specific because it lacks the reference value. Put differently, the plural count noun inherits the properties from the definite article or from article omission. As discussed in Chapter 3, Fried and Östman (2004b: 35) argue that the inheritance relations happening between a determiner and the following noun within the same NP are exclusively unilateral, i.e. from the noun towards the definite article. As their example describes (see section 3.2), the definite article can combine with a count noun (e.g. *the book*) or a mass noun (e.g. *the snow*), inheriting the corresponding features [count] and [mass], respectively. Likewise, since the definite article can occur with both singular and plural nouns, it inherits the feature in relation to the noun number, i.e. [singular] or [plural]. However, it is more plausible that inheritance relations also happen in the reverse direction, namely from the definite article towards the following noun, making the unification process and exchange of properties bilateral. In other words, both the definite article and the empty article slot take on the features of the plural count nouns (i.e. [+plural], [+count]), while the plural count noun receives either the definiteness and specificity information from the definite article (i.e. [+definite], [+specific]) or the non-definite and genericness feature from the omission of the article (i.e. [–definite], [–specific]).

Before proceeding to examine more constructions, it is necessary to make a few amendments to the suggested CxG model. The way the above constructions are represented entails certain limitations and reveals some discrepancies that render it incompatible with the principles of Construction Grammar. The key weakness that needs to be addressed concerns the model's representation of the construction as compositional: the meaning of the constructions in Figure 6.14 is presented as a combination of the semantic properties of their parts. This representation assumes that those semantic properties are fixed, i.e. that for instance the value [+specific] of *the* is fixed for the definite article. As pointed out in section 2.2, the definite article can be used in both specific and generic noun phrases; this feature is thus not fixed but rather dependent on the noun it occurs with. The meaning of these noun phrases is not simply the result of a combination process, but rather, the fillers of the determiner and noun slots interact and together determine the meaning of the whole construction. The way the constructions are designed might therefore contradict the non-

compositionality aspect of Construction Grammar (e.g. Goldberg 1995, 2006). Since the properties of the constructional elements are represented as stable, the suggested CxG model in Figure 6.14 does not include potential phenomena that can occur within and among constructions (i.e. inheritance relations and coercion). Firstly, the model presented in Figure 6.14 fails to properly account for inheritance relations between the article and the nominal head in an NP. The features of a construction's components are not individually added to an NP; on the contrary, a construction as a whole inherits the properties of all its interconnected parts. Secondly, it does not consider the possibility for a noun to be coerced into a construction, i.e. those cases in which a word can change its meaning based on the construction it is alternatively inserted in. Finally, the use of an article may also be predictable: based on what article is used in an NP, the information of the upcoming noun could in turn be predicted. This aspect would therefore be in conflict with another tenet of Construction Grammar, which claims that "[a]ny linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts […]" (Goldberg 2003: 219). Based on the described inadequacies, there is the need to revise and redesign a new constructional model of articles, whose big challenge is also to find a suitable way to represent an element that is not explicitly expressed, i.e. the absence of an article. Neither Fillmore (1988) nor Fried and Östman (2004b), for instance, consider the empty determiner slot found in the generic plural count NP construction. A question that needs to be addressed is therefore whether a feature can be inherited from an element that is not there.

It is from this final point that the revision of the constructional model starts, i.e. the representation of the empty determiner slot and what this entails. As mentioned in section 2.4, Chesterman (1991) and other scholars (e.g. Yoo 2009) claim that bare NPs are in fact preceded by an article, i.e. either the *zero article* or the *null article*. According to Chesterman (1991: 46), the *zero article* is used "for the traditional indefinite article before mass and plural nouns", while the *null article* is used "for the form that occurs before singular proper names." As suggested by Chesterman (1991: 182), articles can be located on a scale of definiteness, where the zero article is the least definite and the null article is the most definite. The indefinite and definite article are positioned in the middle. This definiteness continuum recalls the Prototype Theory of characterization of languages (see e.g. Rosch 1973, Simpson and Miller 1976, and Rosch and Llyod 1978), which presents categories as

prototypes, i.e. as the most representative entity. As stated by Lakoff (1987b: 391), "[d]egrees of category membership for other entities are determined by their degree of similarity to the prototype." Therefore, "the closer an object is to its prototype, the more characteristic it is of the concept" (Osherson and Smith 1981: 37). Likewise, this can be applied to nominals: a nominal is prototypical when it "profiles a physical object instance whose type is specified by a head noun and its grounding by another element such as an article, demonstrative, or quantifier" (Langacker 1999: 23). With respect to articles, their prototypicality relates to the semantic property *definiteness*, whose different degree is determined by the article (i.e. [± definite]).[89] Based on this differentiation, in the constructional model, the zero and the indefinite article result from the feature [–definite], while the definite article and the null article result from the feature [+definite], as shown in Figure 6.15. As Taylor (2003: 220) claims, grammatical categories "have a prototype structure, with central members sharing a range of both syntactic and semantic attributes." According to him, an item still belongs to a category, even if it does not exhibit all attributes. The Prototype Theory will be useful when analysing the elements' properties from a constructional view.[90] That is, similar to the definiteness category, the prototypical structure can be applied to all the other features of the constructional elements.

Based on Chesterman's (1991) classification, both the zero and null articles express meaning. Therefore, in bare NPs constructions, what at first sight seems to be an empty slot contains semantic properties and thus has a meaning. The determiner slot cannot be considered empty as it is filled with either the zero or null article, i.e. by a non-lexical form.

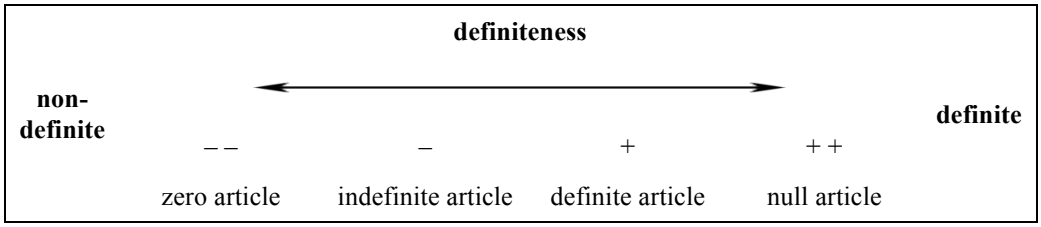

Figure 6.15: Scale of definiteness of English articles.

[89] Specificity is also determined by the articles. For instance, both the definite and indefinite articles can express specificity, but the former is considered more prototypical than the latter. However, while definiteness is a stable feature, specificity is a feature that is realised once the elements within a construction interact.

[90] Note that, for sake of simplicity, the new CxG model keeps the binary system from Fillmore (1988) and Fried and Östman (2004b).

As seen in Chapter 3, constructionist scholars agree that single words can also be constructions, because they combine both form and meaning (see for instance Goldberg 2003: 219). Based on this assumption, it is possible to say that articles can be considered constructions. Talking about article omission is thus not entirely appropriate, because the zero and the null article have meaning and occupy the determiner slot. The particularity with respect to bare NPs is the fact that the form is determined by an element (i.e. either zero or null article) that does not have the "standard" form, i.e. it does not correspond to a lexical form. As explained by Taylor (2003: 243), constructions "are individually learnt as pairings of formal conditions with a semantic specification." In bare NP constructions, the "formal condition" can thus be a component in a covert form. According to Croft (2001: 233), absent elements are not a problem, because "[t]he absence of the overt coded dependency in some contexts does not entail the disappearance of the semantic relation; the semantic relation is simply not overtly coded (it is instead recoverable from other information in the construction or the discourse context)." This therefore strengthens the idea that the illusory "empty determiner slot" in an NP contains semantic properties and thus contributes to the constructional meaning, even if it is not occupied by a lexical item. In the following paragraphs the new CxG model will be presented.

Similar to the first constructional representations, each diagram consists of an outer box that represents an NP and contains two smaller boxes or slots: the one on the left refers to the article (i.e. the determiner slot), while the one on the right refers to the noun with which the article is combined (i.e. the noun slot). The crucial difference between the first and the revised model regards the specificity feature of the article (i.e. [± specific]). Figure 6.16 shows the representation of the adapted CxG model of plural count nouns. As already mentioned, it accounts for the fact that article use is fundamental to the distinction between specific and generic reference of the whole NP.

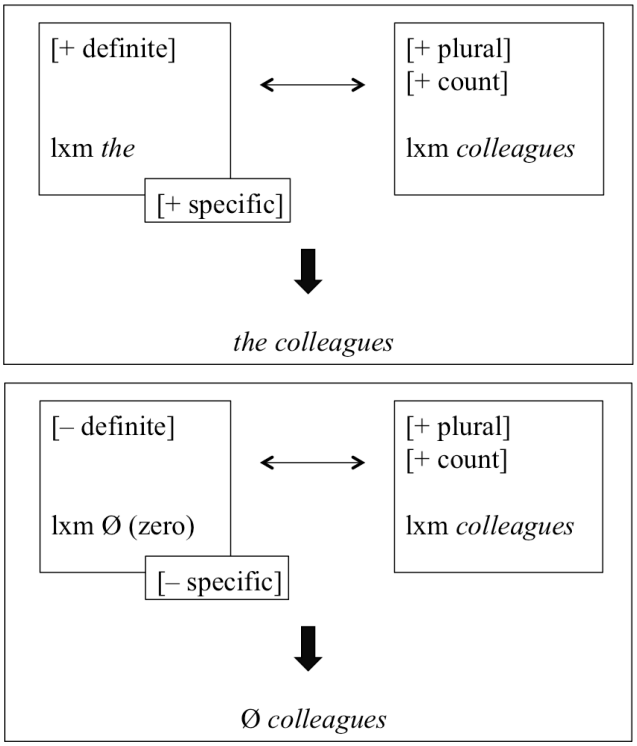

Figure 6.16: Revised CxG model of plural count nouns.

The construction on top in Figure 6.16 represents a plural count noun with specific meaning. The determiner slot contains the definite article and therefore includes the semantic feature [+definite]. On the other hand, the slot for the plural count noun – i.e. *colleagues* – contains the features [+plural] and [+count]. In the new constructional model, the representation highlights the fact that the feature regarding the specificity of the NP is inherited from the definite article and is not a fixed semantic property of *the*. In other words, the meaning of the construction (i.e. a specific plural count NP) is the result of the interaction between the meanings of the definite article and the plural count noun into the meaning of the definite plural count NP construction. Therefore, specificity is not a stable meaning component of the definite article: it is only realised once the article is used in an NP. It is in fact known that the definite article does not always transfer specificity to a definite NP; on the contrary, the meaning component [–specific] can be realised instead, as shown in (1). In the same way, the generic feature can be given by the indefinite article, as in (2), and by the definite article combined with a plural count noun, as in (3).[91]

[91] Examples (1), (2) and (3) are taken from Biber et al. (1999: 265-266). Note that the difference in meaning between the expressions *the Americans*, as in (3), and *Ø Americans* is

(1) The horse is less to the Arab than clay is to the Bursley man.
(2) A doctor is not better than his patient.
(3) The Americans are so jealous because they haven't got a Royal Family of their own.

Articles have therefore meaning *potential* (e.g. specific, unique, familiar), but not all this meaning potential is realised in each construction. The feature that is transferred onto the construction depends on the properties that are unified and exchanged between the parts of the elements included in a construction. Definite plural nouns are a case in point: the meaning component [+unique] is not inherited; likewise, with definite singular nouns, the meaning component [+specific] is not necessarily transferred onto the NP, as shown in (1). Therefore, only by analysing the meaning of the construction as a whole it is possible to determine what meaning is derived (i.e. is inherited) from which article and in which particular context.

The second construction illustrated in Figure 6.16 represents a plural count NP construction with a generic meaning. The properties included in the slot of the plural count noun remain the same (i.e. [+plural] and [+count]). What differ are the features of the determiner slot. As discussed before, in a bare plural count NP the zero article is used and its semantic feature is thus [–definite]. The meaning of the zero article combined with the meaning of a plural count noun results in a generic NP. This is represented by the property [–specific], which is invoked by the zero article and transferred onto the whole construction. Contrary to Fried and Östman's (2004b: 35) statements, as already argued in the previous section, the inheritance relations between the construction's two components happen from both directions. In other words, both parts have an active role within the construction and contribute to the construction meaning.

Figure 6.17 shows the structure for a singular proper noun. The semantic features included in the noun slot are [+singular] and [+proper]. The determiner slot is filled in by the null article, which, according to the scale of definiteness suggested by Chesterman (1991: 182), represents the article with the highest degree of definiteness. Its semantic property is therefore [+definite]. This is an interesting case that differs

very subtle and might depend on external factors. In the former, the definite article is used to single out *Americans* as a group, comparing them with other groups, e.g. *The Americans are coming!* can be used in a context in which different types of *people* are defined based on their nationality; in the latter, with the zero article they are seen as an open class. Even though both expressions denote a generic meaning, the definite article is less generic than the zero article.

from the others because, as discussed by Fillmore (1988: 40) and Fried and Östman (2004b: 40), proper nouns are already maximally specific (i.e. they are naturally specific). In other words, they naturally contain the specificity property, and the definite article in turn cannot occur. If this happened, the unification process would not work successfully, as the constituents' information would be in conflict, because the same feature would be represented twice.[92] Proper nouns dominate this construction and demand the presence of this type of article – i.e. the null article – because it complies with their specificity. Put differently, specificity is maximalized by the proper noun and therefore prompts the presence of the null article, which is compatible with this construction.

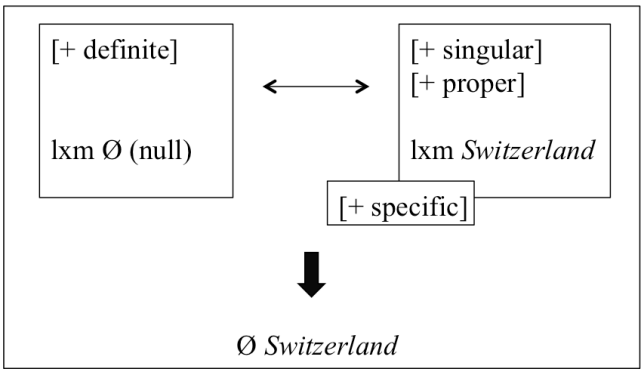

Figure 6.17: Revised CxG model of singular proper nouns.

Hence, it is possible that, instead of adding the specificity value to the construction as it usually happens in regular NPs, the function of the null article here is to match the specificity feature already provided by the proper noun. As shown in the construction diagram, the value [+specific] is inherited from the proper noun and transferred onto the NP, i.e. the micro-construction *Ø Switzerland*.[93]

[92] Note that it is sometimes possible to use a definite article with a proper noun. Consider for instance the following sentences: *This is the Italy I love* or *The Paul we saw on stage was brilliant*. In these cases, the unification of a proper noun with the definite article is possible due to inheritance relations that connect the values of the NP construction with the values of other constructions, which are external of the noun phrase.

[93] If the indefinite article were used instead, the non-specificity feature would derive from the article because it would override the specific value of the proper noun. In fact, the use of the indefinite article with proper nouns is very marked (e.g. *She looks like a Jennifer* creates a large set of "Jennifers", or *They want to build a European Union* makes "European Union" a potential rather than an existing entity. The latter could happen in the context of a European Union's reform, whose aim could for example be to create an European Union in which Member States have more equal rights.

Singular abstract nouns are modelled in Figure 6.18 – Figure 6.20. As seen in section 6.3, abstract nouns mainly occur as bare NPs. The underlying constructional representation is given in Figure 6.18. In this construction, the box including the properties of the noun has the features [+abstract] and [–proper], while the slot of the determiner is filled in by the zero article and thus comprises the feature [–definite]. In order to reach the generic meaning realised in this construction, the property [–specific] is inherited from the zero article, while the feature [–count] is inherited from the noun. Again, these features are transferred onto the NP as the result of the interaction between the components of the construction. More specifically, the non-count abstract NP construction inherits the non-countability value from the abstract noun – i.e. *legislation* – and inherits the generic reference from the zero article. In this instance, the example *legislation* therefore refers to the process of making laws, rather than a specific law.

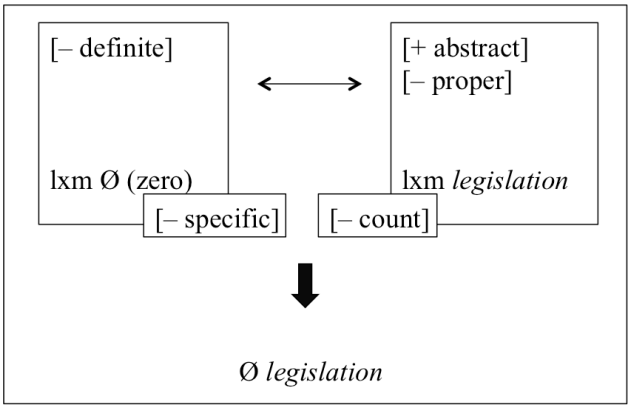

Figure 6.18: Revised CxG model of uncountable abstract nouns.

Abstract nouns are of particular interest because there are different kinds (Quirk et al. 1985: 247): they are prototypically non-count, but some can also be count (e.g. *legislation, law*), while others are only non-count (e.g. *progress, information*). Figure 6.18 can therefore be considered as the prototypical structure of an uncountable abstract noun. However, as shown in Figure 6.12, non-count abstract nouns sometimes occur with the definite article.[94] The underlying construction is illustrated

[94] Note that the use of the indefinite article with a non-count noun is possible when the noun is premodified. Compare e.g. *She played the oboe with a* sensitivity* and *She played the oboe with (a) charming sensitivity* (Quirk et al. 1985: 287). Similar to the possibility for a proper noun to be combined with an article, the use of the indefinite article with a non-count noun

in Figure 6.19. The upper structure refers to an abstract noun that can be either non-count or count (e.g. *the legislation*), whereas the structure below refers to an abstract noun that can only be non-count (e.g. *the progress*). In both structures, the determiner slot contains the property [+definite], and the definite non-count abstract NP construction thus inherits the feature [+specific] from it; on the other hand, the noun slot includes the values [+abstract] and [–proper]. What differs between the former construction and the latter is the position of the feature [–count]: in the first one, the property [–count] is outside the noun slot, as it is realised in this construction, while in the second one, the same feature is a constant property and is therefore located inside the noun slot, as the abstract noun in question is only uncountable.[95]

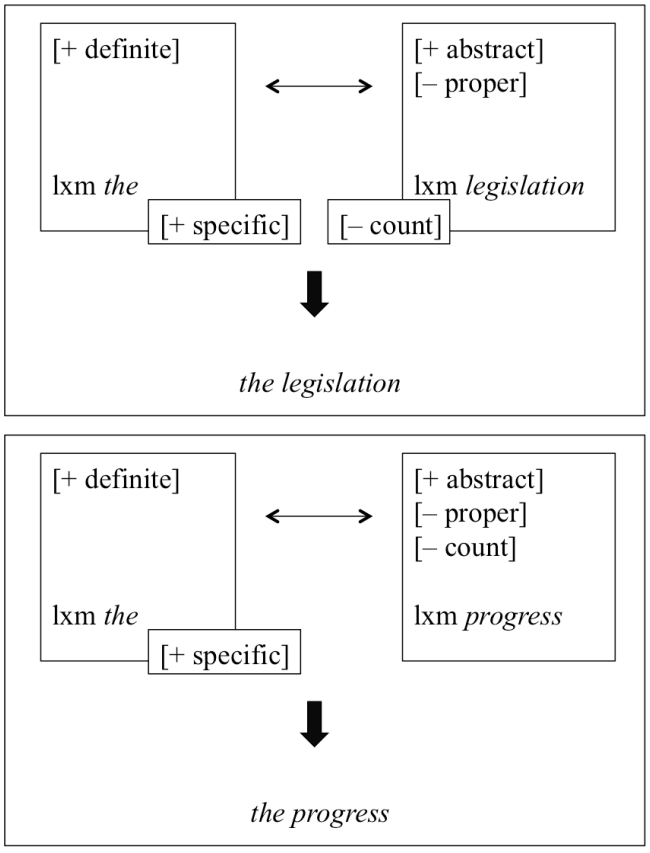

Figure 6.19: Revised CxG model of uncountable abstract nouns with the definite article.

creates a larger set of "sensitivity types" that differ from each other based on a quality that given by a premodifying element, i.e. an adjective.

[95] Note that the construction of *the progress* can clearly occur with the zero article (i.e. *Ø progress*) and inherit the feature [–specific], analogous to Figure 6.18.

Figure 6.20 shows the constructions of an abstract noun preceded by an article and behaving like a countable noun. The first diagram shows a definite countable abstract NP construction. The determiner slot includes the feature [+definite], while the properties of the noun that follows are [+abstract] and [–proper]. A possible interpretation[96] for this instance is that the interaction does not occur within the same construction but between two different constructions. One is the definite noun phrase construction, which prototypically consists of the definite article and a singular concrete count noun (e.g. *the candle*), the second is the non-count abstract noun phrase construction (see Figure 6.18), which consists of the zero article and an uncountable abstract noun. Since an abstract noun is prototypically non-count, when it is inserted into the constructional frame of the definite noun phrase construction, the abstract noun is coerced into a count-noun reading. The unification process between these two constructions (i.e. the definite article combined with a singular concrete count noun, and the zero article combined with an abstract non-count noun) results in a countable abstract noun with specific reference (i.e. the definite article combined with an abstract count noun), in which the feature [+specific] is inherited from the definite article, and the feature [+count] is the result of the coercion effect. As discussed in the section 3.1, the principle of coercion applies when the meaning of a lexical item varies with the contexts of the construction it is used in. As Michaelis (2004: 25) explains, "[i]f a lexical item is semantically incompatible with its morphosyntactic context, the meaning of the lexical item conforms to the meaning of the structure in which it is embedded." The interaction between the two parts thus makes the noun *legislation* (primarily uncountable) a countable noun; additionally, contrary to the construction shown in Figure 6.18, the micro-construction *the legislation* then has a different meaning, namely the result of the act of creating laws (i.e. the law passed by the government).[97]

[96] Note that this interpretation is merely speculative. In order to be able to compare constructions and to confirm the suggested reading, another sample containing a larger variety of nouns should be retrieved.

[97] It could be argued that the countable sense of *legislation* might not be abstract but rather concrete because *the/a legislation* could have a physical attribute (e.g. a law typed on a sheet of paper). However, it is considered here as an abstract noun because it is semantically close to the concept of *principle*, which is considered an abstract noun (Crystal 2008: 3).

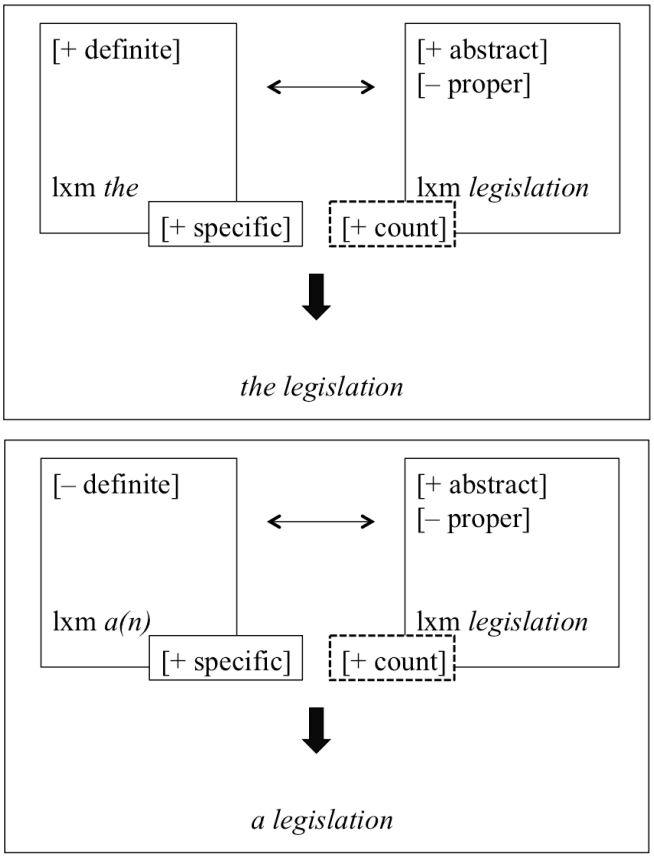

Figure 6.20: Revised CxG model of countable abstract nouns (coercion effect).

The second structure shows an indefinite countable abstract NP construction. The features of the noun do not change (i.e. [+abstract], and [–proper]), while the property in the determiner slot is [–definite]. Similar to the previous construction, the resulting construction meaning is an abstract and countable noun with specific reference. The feature [+specific] derives from the article, while the property [+count] is the result of the coerced abstract noun and is transferred onto the whole construction. As discussed in the prior representations, since abstract nouns are prototypically uncountable, these constructions represent the result of a coercion effect, because abstract nouns become countable when inserted in a definite count noun construction (i.e. the construction thus occurs between a construction and a noun). The coercion phenomenon discussed here strengthens the notion that constructions form a constructional network – i.e. a construct-i-con – in which constructions are connected to each other via various links, at the same (or different) level(s) of abstraction (Hilpert 2014: 63). The different types of interactions between constructions within the constructional network thus prove that the resulting meaning of an NP cannot be simply predicted by the meanings of

the construction's components. Also, the meaning of the construction cannot be inferred by the meanings of its parts; rather, it is derived from the interaction between the meanings of the (non-)lexical items resulting into the final meaning of the construction (i.e. via inheritance relations or coercion phenomena). It cannot, therefore, be considered compositional.

So far, this section has discussed the constructional properties of NPs that are not postmodified by an *of*-phrase. The following part continues with the analysis of abstract nouns within an *of*–CONSTRUCTION. As mentioned before, based on the construction hierarchical system suggested by Traugott (2008) and Trousdale (2008), the *of*–CONSTRUCTION is analysed here as a meso-level construction, i.e. a more defined construction of a nominal postmodified by a PP, which is located at the macro-level. Figure 6.21 shows the constructional representation of an abstract noun followed by an *of*-phrase and preceded by the definite article. In the graph, the outer box (i.e. the construction) includes three boxes: one for the definite article, one for the abstract noun, and one for the *of*-phrase. The determiner slot includes the property [+definite], and the abstract noun slot contains the properties [+abstract] and [–proper]. The whole construction inherits the feature [+specific] from the definite article and the feature [–count] from the abstract noun. On the other hand, the *of*–phrase is defined as [+distinctive] because it gives more information about the noun it is combined with. The use of an *of*–CONSTRUCTION thus confines the attention on a restricted aspect of the noun; for instance, in the expression *the loss of life*, the preposition *of* makes the whole NP limited to *life* and distinguishes it from other types of losses, e.g. *the loss of money* or *the loss of dignity*. Therefore, the property that is inherited from the prepositional phrase and is transferred onto the final construction meaning is defined here as [+narrowness].[98]

[98] Note that the feature [+narrowness] also includes possession meanings. For instance, in the case *the tail of the dog* the preposition *of* expresses a relation of possession or ownership and still narrows it down, i.e. it refers to *the dog's tail* and not *the cat's tail*. The choice of terminology aligns with Croft (1991: 52) and Taylor (2002: 352), who speak of a narrowing effect when discussing nominals' modification.

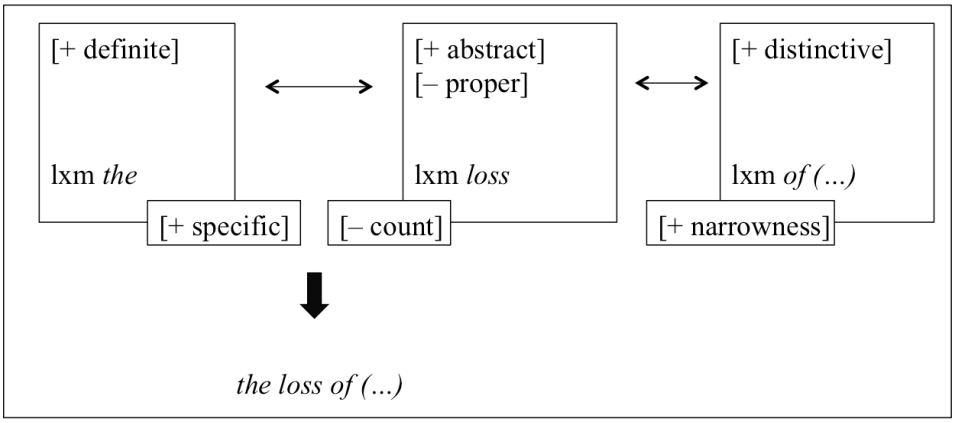

Figure 6.21: Constructional representation of *of*–CONSTRUCTION with specific reference.

The feature of narrowness given by the *of*-phrase to the construction thus prompts the high level of specificity with the definite article; put differently, when these two elements are combined, they express high degree of specificity. This phenomenon therefore explains the frequent occurrence of an abstract noun postmodified by an *of*-phrase with the definite article.

Figure 6.22, on the other hand, shows the representation of *Ø loss of (...)*. The diagram models the construction largely in analogy with the previous one. The only difference appears in the slot of the article. The determiner slot is filled with the zero article (i.e. [–definite]), which in turn adds the generic reference to the construction (i.e. [–specific]).

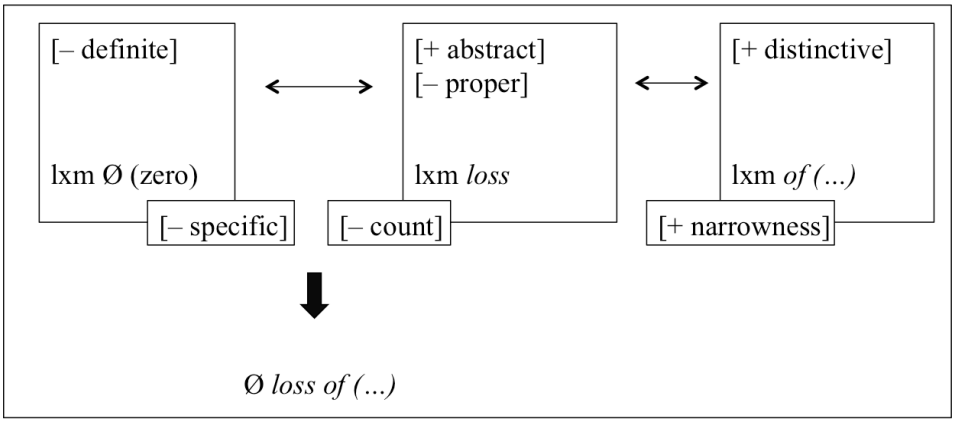

Figure 6.22: Constructional representation of *of*–CONSTRUCTION with generic reference.

In NPs modified by an *of*–phrase, the set of instances that the modified head noun can refer to is narrowed, i.e. the *of*-phrase delimits the scope of reference of the head noun. In conjunction with a definite article, this narrow reference is then made specific. In contrast, when the determiner slot is occupied by the zero article, the whole construction is given a generic meaning, but the narrowness feature remains.

In general, every construction conveys a distinct meaning. In the contexts that have been described, article use is essential to the meaning of the NP: the specific reference is realised by the definite, indefinite and null article, while the generic reference is realised by the zero article.

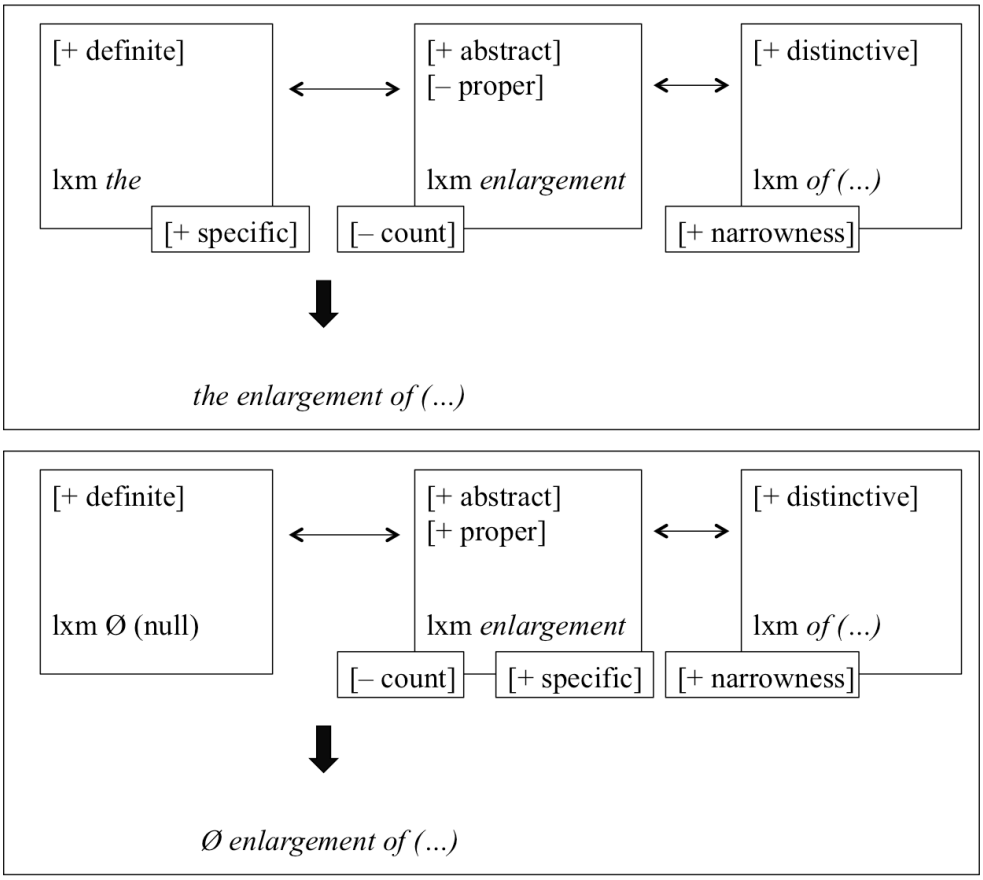

Figure 6.23: Constructional representations of article variability.

The final aspect that needs to be pointed out regards the case of free variability (i.e. variation without change in meaning). In the follow-up analysis, the only discovered case of variable article use was *the/Ø enlargement of the European Union*. The corresponding representations of *the enlargement of (...)* and *Ø enlargement of (...)* is shown in Figure 6.23. The construction meaning (i.e. full specificity) is shared by

both expressions, i.e. they are two independent and synonymous variants. The meaning, however, derives from different types of interaction between the elements, i.e. from different semantic properties. On the one hand, the high level of specificity contained in the whole construction is the result of the interaction between the definite article, an abstract noun, and an *of*–phrase. On the other, full specificity is inherited from the interaction between the null article, an abstract noun, and an *of*–phrase. What changes between these variants is the feature of the abstract noun, which is either proper or non-proper (i.e. [±proper]). In the former variant, the noun *enlargement* follows the definite article and is not proper (i.e. [–proper]). In the latter, as already discussed in section 6.4, it appears as bare because it might behave more like a proper noun (i.e. [+proper]), which in turn combines with the null article.[99] As a whole, the resulting construction meaning is determined by a high degree of specificity, which is given by a different kind of inheritance relations. The semantic shift attested in the bare construction (i.e. from [–proper] to [+proper]) may therefore arise from factors that are external to the NP, such as the context;[100] namely, the meaning that *Ø enlargement of the European Union* has within the context of the European Parliament and among its members. With Construction Grammar, therefore, it is possible to clarify those cases that do not follow the general patterns. Even though their frequency is very low in a corpus, they can be explained from a constructional perspective (Goldberg 2013: 17).

In the case of article use, the last big challenge is to show how all the above constructions are actually connected to each other. The main question is whether there are perhaps some constructions that should be seen as descendant of or are derived from another construction (i.e. vertical relation) or whether they all need to be analysed on the same level (i.e. horizontal relation). Figure 6.24 is the visualization of the generalization that can be obtained by connecting all the constructions that have been analysed and discussed in the current chapter. More specifically, the individual examples of corpus evidence that led to the above constructions are given below each

[99] While treating *Ø enlargement of the European Union* as a generic NP may seem intuitive, it would directly undermine the variability of the structure. For most corpus examples, it is difficult to make a case that there is a difference in meaning with or without the preceding article.

[100] Note that the diagram for the bare NP variant is not fully representative, as the feature [+proper], in this case, is not stable but derives from external factors (i.e. context). Therefore, this property could alternatively be located on the upper part of the noun slot.

micro-level representation for better understanding. Also discussed examples from Chapter 5 are presented in the model as additional evidence (in colour).

The analysis moves from a higher level of abstraction, i.e. the noun phrase construction,[101] to the meso-level, i.e. abstractions immediately preceding the more detailed constructional schemata that were presented and discussed in detail in the current study. Boxes are used again to represent the constructions. In order to facilitate reading, higher-level constructions are not fully specified but referred to with labels only. The arrows symbolize the horizontal relations of functions without repeating preceding information. Specifications of the constructions are therefore given at the bottom of the diagram, in which the different constituents (needed to build the overall meaning of the constructions) are included. The elements contained in curly brackets are those optional components that may appear within a noun phrase and contribute to the meaning of the construction. What is important to point out is that, once again, this final CxG modelling is strictly based on empirical evidence and limited to constructions within the NP. However, it also extends the discussion by referring back to the individual findings described in Chapter 5. For sake of completeness, the diagram thus takes into consideration the most relevant results discussed in the analysis of (variable) article use with collectives, namely the importance of premodification within an NP (marked in green in Figure 6.24).

[101] For illustrative purposes, the highest level of abstraction in Figure 6.24 connecting the various constructions is just labeled as NP, which according to Table 6.7 would be the macro-level. Given that this study only considers NPs with articles (and not other determiners), this highest level actually represents what in Table 6.7 is meso-level (2).

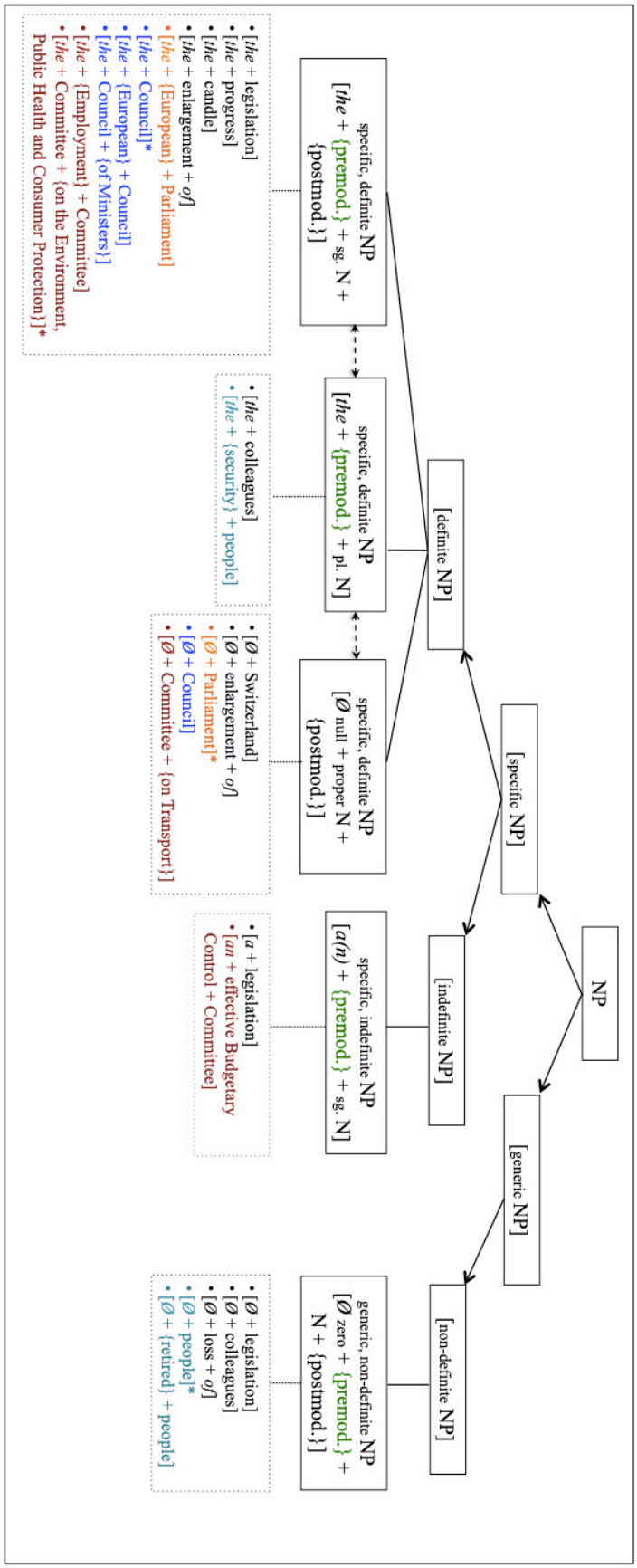

Figure 6.24: Data-based outline of the constructional network of article use down to meso-level.

The noun phrase is at the top of the constructional network (i.e. the highest level of abstraction). By following the hierarchical organization, on a lower and more detailed level, a first distinction needs to be made, namely whether the elements included within the NP, as a whole, refer to a specific noun phrase (i.e. [specific NP]) or a generic noun phrase (i.e. [generic NP]). From the retrieved and analysed data, it was possible to characterize a more specified construction derived from the generic NP, that is a non-definite (generic) NP (i.e. [non-definite NP]).[102] More specifically, this refers to a construction that contains the zero article and a noun. The latter refers to an unspecified noun, which means that it can be either singular or plural, concrete or abstract, countable or uncountable (as seen in the constructional schemata of the constructs *Ø legislation* and *Ø colleagues*). Furthermore, as previously discussed, the construction can include a postmodification slot, i.e. the *of*-phrase (e.g. *Ø loss of*). By looking at the analysis of article use in relation to collective nouns, it is possible to see that *people* perfectly falls in this construction, as it refers to a plural noun with generic reference that can occasionally be premodified by an adjective (i.e. *Ø {retired} people*).

With respect to the specific NP, a further differentiation is required. A specific noun phrase can in fact be either definite or indefinite. In the case of indefiniteness, an NP always includes an indefinite article and a singular noun, which can be concrete or abstract, as previously discussed in the example *a legislation*. The case study related to the collective nouns revealed that a premodification slot can be added to this construction as well, as shown in the example of *an effective Budgetary Control Committee*.

The function expressed by a definite noun phrase, derived from a specific NP, can be made explicit through several forms, i.e. via three different constructions. One construction contains the definite article and a singular (common) noun, which can be concrete (as assumed in the prototypical case of *the candle* and discussed with the example of *the Council*) or abstract and uncountable (e.g. *the legislation/progress*). Furthermore, this construction allows for a postmodification constituent with the *of*-phrase, as found in the case of *the enlargement of*. In the study of the collective nouns, premodifying elements were found to play a significant role in the specificity

[102] As previously mentioned and shown by Biber et al. (1999), also an NP containing a definite or an indefinite article can derive from a generic NP, but these variants are not presented here as they are not supported by data from this study.

aspect of a (definite) noun phrase, as the examples *the European Parliament* and *the {European} Council* showed.[103] Therefore, an optional slot containing a premodifying element needs to be included in the construction.

A further possible construction having specific and definite meaning contains the definite article and a plural noun (shown in the construct *the colleagues*). Like in the previous construction, a slot referring to premodification can also be inserted. An example is again the plural collective noun *people*, which occurred with premodifying elements, such as *security* in the construct *the security people*.[104]

The third and last construction referring to a specific and definite NP consists of the null article combined with a noun that is naturally proper, such as *Ø Switzerland*, or a noun that is (or might be) interpreted as proper, such as *Ø Parliament* and *Ø Council*. This construction has also the option of a postmodifying element, as shown in the case of *Ø enlargement of*, in which the noun *enlargement* behaves like a proper noun.[105]

These three constructions are of particular interest, as they show that NPs defined by the semantic features [+definite] and [+specific] can be realized differently. This means that NPs with similar meaning (i.e. a definite noun phrase containing the specificity component) have, on the formal level, three different realizations.

So far, the analysis has followed the conception that constructions need to be described separately and independently (see e.g. Goldberg 2002). With respect to article variability, however, the above constructions have shown that the specificity or genericness aspect of an NP can be realised through different complex compositions. A specific, or generic, noun phrase can therefore be achieved by choosing from a set of possible constructions. Put differently, in order to make an NP either specific or generic, regular *alternations* are available for the speaker and can be used. This is particularly evident with the three constructions referring to a definite and specific NP. Construction Grammar, however, generally argues against the idea of alternations

[103] Also additional postmodifying examples occurred, corroborating the already discussed postmodification slot, e.g. *the Council of Ministers*.

[104] Cases of specific and definite NPs containing a plural noun with postmodifying elements were not found in the data sets discussed in the current study and are therefore not included in the representation of the constructional network. Otherwise a unification of this construction with the one presented to the left could be argued for.

[105] Also for this construction, optional premodification could probably occur (e.g. *Ø beautiful Italy*) but is not included, as there is no supporting evidence in the underlying data.

stored in the speaker's linguistic inventory. Goldberg (2002), for instance, argues that the *spray/load* "locative" patterns (e.g. *spray the wall with paint* and *load the wagon with hay*) simply need to be seen as separate and independent constructions and not alternations to the caused-motion pattern (e.g. *spray paint on the wall*, *load hay onto the wagon*). They are thus seen as a transformation of a construction, not a variation of it. As Goldberg (2002: 329) states "[t]here are typically broader syntactic and semantic generalizations associated with a surface form than exist between the same surface form and a distinct form that it is hypothesized to be syntactically or semantically derived from." However, the constructional network represented in Figure 6.24 clearly shows that the constructions are not only connected vertically (i.e. by inheriting the semantic features derived from higher levels of abstraction), but that they can also be connected horizontally (i.e. on the same level). As mentioned before, this is easily visible when looking at the three constructions that refer to specific and definite NPs: they have the same meaning, which is expressed through different forms. They can be treated as alternations and not as distinct and separate constructions. Alternation-based generalizations were presented and discussed by Cappelle (2006), who suggested describing closely related constructions not as derived from each other but as *allostructions*. These are defined as "variant structural realizations of a construction that is left partially underspecified" (Cappelle 2006: 18). Similar to morphology having different formal realizations of the same morphological unit (i.e. allomorphs), other parts of grammar (e.g. phrasal units) can be expected to have similar variations on the formal level (Cappelle 2006: 21). Alternation-based generalizations (i.e. allostructions) thus refer to generalizations that "are based on semantic similarities between formally distinct constructions and capture the fact that a given event type may be expressed in various ways" (Perek 2012: 608). In the case of article variability, the event type in question relates to the [±specific] semantic property expressed by the combination of the elements included in the noun phrase. Cappelle (2006) uses corpus data to identify the more frequent of two alternating allostructions, which he then marks as the default version. Given that the constructions presented here also have underlying corpus data, it is possible to identify the more frequent allostructure of two variants. For example, in Chapter 5 the case study showed that *Council* appears most frequently with the definite article and without premodification, therefore making [*the Council*] the favoured allostruction over [*the European Council*] or [*Ø Council*]. This preference for one allostruction

over the other is therefore not fixed at the meso-level but depending on choices at the micro-level.

A possible conclusion from an interpretation of the construct-i-con with allostructions is that article use can be seen as variable for certain constructions indeed, i.e. those that are confined within the scope of a shared allostruction template and therefore share the function, e.g. *the/Ø enlargement of* as specific definite NP. As counter example, *the/Ø colleagues* cannot be considered truly variable, as the specificity function differs.

6.6 Beyond the NP

Although the analytic focus of this thesis lies within the scope of the NP, it became clear that some of the examples discussed above cannot be fully accounted for by what is happening within the NP. Therefore, the question remains how the proposed CxG model interacts with structures that the NP is embedded in. Given the many types of links that exist in a constructional network and the many variables at the sentence level that could potentially influence the choices at the NP level, this question can of course not be answered with a single "one template" explanation but rather needs to be looked at case by case. Interesting examples are (8), i.e. *present to Ø Council*, and (9), i.e. *ask Ø Council*, from Chapter 5 regarding the article use of the collective noun, which usually appears with an article but given these examples seems variable. In the model shown in Figure 6.24, it can be seen that *Council* as a specific, definite NP mostly occurs with a definite article, but, given the available allostructions, a pattern without the definite article, and which is thus more like a proper noun (PN), is also possible. This means that, without the external influence beyond the NP, *Council* tends to be used mainly with a definite article. In the case of example (9) from Chapter 5, however, the NP is viewed in the context of a verb phrase, i.e. VP (*I ask […]*). The act of asking normally requires an animate object. If an institution fills the slot of the animate object, a PN-like usage of the institution becomes an option, as already discussed with *Parliament* in Chapter 5. Therefore, it can be argued that the NP object in *I ask Ø Council* takes on PN-like properties, thus evoking the PN allostruction option of the specific, definite NP *{the|Ø} Council* – not as a preferred choice, i.e. more frequent occurrence, but as a grammatical possibility, the less frequent allostruction, in this case.

Example (8) in Chapter 5 leaves more room for interpretation. On the one hand, it could simply be considered a case of Goldberg's (2013) P N construction discussed in section 3.1, i.e. a combination of a preposition with a bare count noun such as *to bed*, but that would limit it to cases with directional or locational meaning. On the other hand, the same logic of multiple inheritance from above can be applied to example (8) specifically, where the verb *to present (to)* needs an animate (indirect) object, making the PN allostruction of the NP *{the|Ø} Council* a speaker's possible choice. Thus, the property of the external constructions (in this case, the required animate property of the verb's object) is the linguistic context that influences the choice of the particular variant among the allostructions of the specific, definite NP. It can be argued that this process is replaced by an independent construction if the specific combination of words occurs frequently enough to postulate a construction in its own right. From what moment exactly this is the case is an intriguing question for further research.

Figure 6.25 shows a representation of this multi-inheritance link where the micro-level NP *{the|Ø} Council* independently favours the allostruction with the definite article, but interaction with the external context of the VP (e.g. *ask {the|Ø} Council* or *present to {the|Ø} Council*) extends the options to the allostruction without article. The preferred allostruction is marked with bold lines and the less frequent choice marked with dotted lines.

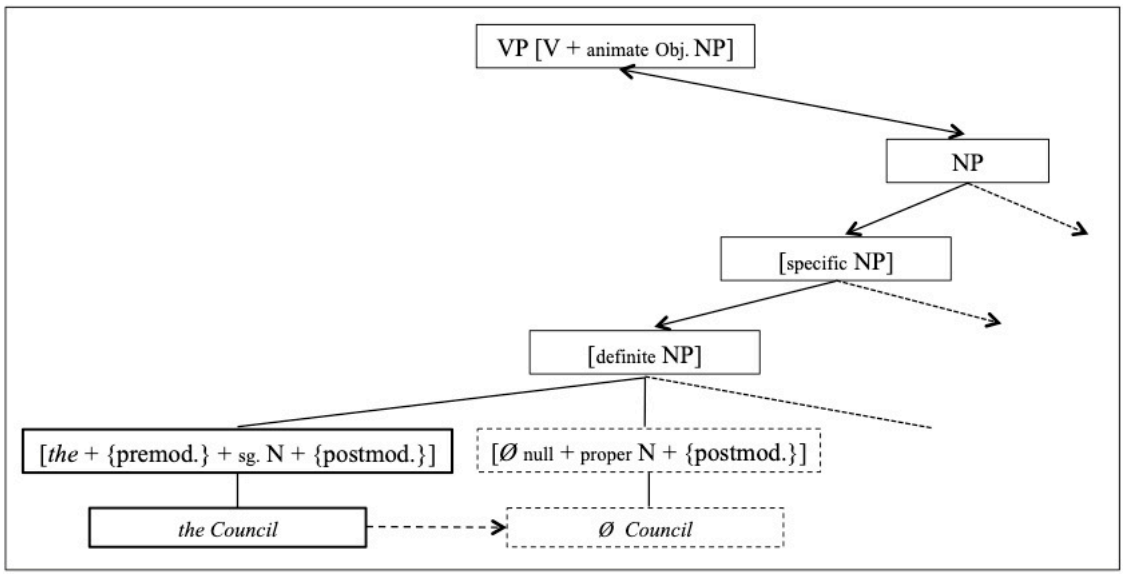

Figure 6.25: Representation of V influence on NP allostructions.

A similar case can be made for phrases that follow idiomatic patterns, such as example (57), i.e. *at Ø European level*, and (58), i.e. *at Ø local level*, from section 6.4. The premodified NP *the* ADJ *level* in isolation can be placed in the specific, definite NP slot on the presented constructional network model shown in Figure 6.24. But again, if it merges with an external construction, this can influence or, in this case, override its natural property. A case can be made for an existing construction [*at* + height indication] that never appears with an article, such as *at Ø 300m*, *at Ø sea level*, *at Ø eye level*. It appears that certain lexeme-combination with *level* metaphorically behave the same way, i.e. by extending the physical height indication to a more abstract hierarchical height, as seen in the examples presented in (57) and (58). Thus, cases such as *local level* can be found with or without the article, depending on whether the speaker is guided by the metaphorical template of the fixed expression or by the grammatical default option of the NP itself. The metaphorical extension of a fixed construction such as the [*at* + hierarchy indication] construction derived from the [*at* + height indication] construction – via metaphorical link – would thus override the default NP dynamics as presented in Figure 6.26.

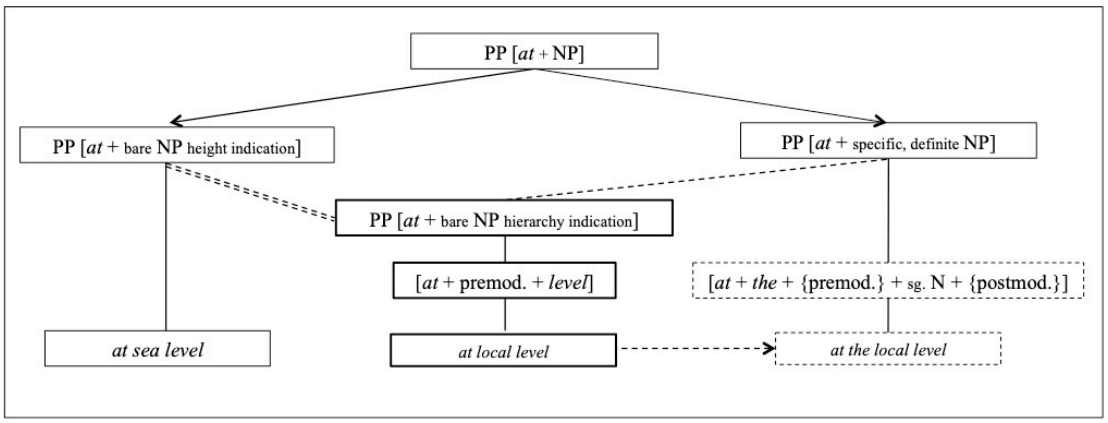

Figure 6.26: Representation of *at + hierarchy indication construction* with its construct-i-con links.

In a way, thus, the metaphorical construction [*at* + hierarchy indication] becomes an allostruction itself, in the sense that the speaker is faced by variability without change in meaning and arguably guided by frequency. Therefore, the same style of representation as established for allostructions has been chosen for the model in

Figure 6.26, where the new, preferred allostruction is given in bold and the metaphorical link is represented with a double-dashed line.

6.7 Summary

The main aim of this investigation was to retrieve a dataset containing English bare NPs with a data-driven approach. The retrieval process entailed various steps, which were required to both define general characteristics of bare noun phrases and to build a constructional model for the analysis of (variable) article use. Taken together, the results supported what standard grammars claim, namely that bare NPs in English mainly occur with non-count nouns and plural count nouns (see e.g. Quirk et al. 1985: 282) and convey a generic reference (see e.g. Biber et al. 1999: 265). The final dataset including only bare NPs showed a high frequency of abstract nouns (over 65%). This is due to the methodology adopted in this study, i.e. the use of German definite NPs for the retrieval of aligned English bare NPs. Finally, the findings showed that articles are more variable among plural nouns (75%). Articles are used here to clearly distinguish between specific reference (i.e. article use) and generic reference (i.e. no article). On the other hand, in the singular subset, variable contexts were found with those abstract nouns that are primarily uncountable but that can also have a countable meaning. The way an article is used generally determines the meaning of the noun in question: they are usually bare when they are non-count, while they take an article when they are count (e.g. *legislation* indicates an act of making laws, while *the legislation* can denote a law established by a legislative body).

The follow-up analysis narrowed down the data selection and paid particular attention to abstract nouns, i.e. the comparison of abstract nouns without postmodification and abstract nouns postmodified by an *of*-phrase (i.e. the *of*–CONSTRUCTION). The dataset was further subdivided according to the presence of the *of*-phrase as well as the article. Corpus evidence confirmed the tendency for abstract nouns to mainly occur as bare NPs but to take an article when postmodified by an *of*-phrase (Quirk et al. 1985). However, the data also showed that the construction in which a bare abstract noun is postmodified by an *of*-phrase is definitely possible. Several chi-square tests were applied to investigate whether the factors 'premodification', 'countability' and 'reference' influenced article use. The influence of countability and reference was significant. Hence, articles normally occur when an

NP contains a countable abstract noun and when it has specific reference. In the dataset, one particular instance of free article variability was found in the *of*–CONSTRUCTION, i.e. *the/Ø enlargement of the European Union*. A plausible explanation for the article to occur variably in this noun phrase is that, in the parliamentary context, the expression *Ø enlargement of the European Union* is not seen as a free phrase, but rather as one unit. Therefore, the head noun *enlargement* behaves more similarly to a proper noun, rather than a regular abstract noun.

The last part of the analysis focused on the application of the Construction Grammar framework in relation to (variable) article use. The data-driven approach provided a strong empirical basis for the constructional analysis, which thus entailed a bottom-up perspective. The presented constructional model took into account the phrasal level and, for sake of simplicity, it was strictly based on a simple NP version, i.e. premodification was not included. Furthermore, since the focus was on the semantic properties of the elements within an NP and the semantic properties transferred onto the construction as a whole, the syntactic functions and external factors (e.g. context or discourse) were not taken into account. The analysis looked at the constructions relating to plural count nouns, proper nouns, abstract nouns (including those that can be both non-count and count and those that are only non-count), and abstract nouns postmodified by a PP, i.e. the *of*-phrase. The first constructional model was based on the Determination Construction proposed by Fillmore (1988) and Fried and Östman (2004b). However, since their versions did not fully respect two relevant principles of Construction Grammar (i.e. the non-compositionality and non-predictability aspects of constructions) and did not properly account for potential phenomena that can occur within and among constructions (i.e. inheritance relations and coercion), a revised CxG model was needed. The new CxG model tried to fulfil the above-mentioned limitations and aimed to better describe the inheritance relations among the elements within a construction, which, contrary to Fried and Östman's (2004b) claim, do not only happen from the noun towards the article, but also from the article towards the noun it is combined with.

Another relevant aspect that the revised constructional model focused on was the analysis of the empty determiner slot of a bare NP. It was shown that all constructions convey a defined meaning; for instance, the difference in meaning between the two structures regarding a plural count noun lay on the reference: the construction occurring with the definite article results in a specific NP construction,

while the one occurring as bare noun phrase results in a generic NP construction. Therefore, it is likely that the "empty" determiner slot shares properties with the following noun and contributes to the meaning of the construction. It was then argued that the determiner slot is always filled with an article. More specifically, based on the classification of articles suggested by Chesterman (1991), the null article occupies the determiner slot when the nominal head is a proper name, while the zero article fills it in front of the other possible bare nouns (e.g. plural nouns, mass nouns, abstract nouns). All articles thus contain the semantic feature of definiteness, which is however expressed to different degrees. Hence, based on this conception, talking about article omission is not the most ideal interpretation. Taken together, therefore, the analysis showed that articles are constructions themselves, as they are pairings of form and meaning. The particular aspect of the zero and the null article is that they convey meaning without a standard form (i.e. lexical form). Thus, the determiner slot still entails semantic relations, which are not overtly expressed (Croft 2001: 233).

Moreover, it was argued that NPs are therefore the results of an interaction of constructions, i.e. the article construction and the noun construction. Put differently, constructions are determined by a countless number of nodes (i.e. interactions) resulting in a large variety of meanings, which strictly depend on the elements' properties and the relations between the constructional elements. As shown in the various diagrams, the constructions share features and interact differently, resulting in various different meanings (compare e.g. the construction of a proper noun, in Figure 6.17, with the construction of a non-count abstract noun, in Figure 6.18). An interesting case was the interaction between a regular count concrete NP construction with a non-count abstract NP construction, in which a non-count abstract noun is inserted in a different construction frame resulting in a coercion phenomenon (i.e. the abstract noun is coerced into a count noun). Constructions constantly interact with each other and are linked in various ways. This counts for constructions that are on the same level of abstraction, but also on a different level. A further interesting instance regarded the expression *the/Ø enlargement of the European Union*, in which two independent constructions share the same meaning (i.e. full specificity) but are differently structured. From a constructional perspective, it was interesting to explore what relations occur within the NP and among the elements. The crucial difference is determined by the noun *enlargement*: when preceded by the definite article, it behaves like a prototypical abstract noun postmodified by an *of*-phrase, while it behaves more

like a proper noun when appearing as bare NP, i.e. occurring with the null article. This instance might be referred to as a case of free variability; however, the addressed question was whether it can be considered as such, as the specificity value transferred onto both constructions derives from two different positions. In order to further address this question, an outline for a constructional network was presented in Figure 6.24, proposing a viewpoint that the example of *the/Ø enlargement of* can be considered as variable indeed. By suggesting the occurrence of allostructional relations among some of the presented construction frames, it can be explained why some cases are truly variable, while others undergo a change in meaning and therefore vary only at the surface. Finally, a look beyond the NP revealed that there are indeed factors outside the NP that seem to lead to variability in the choice of article use. Although this will need further in-depth research to verify, a case can be made for both fixed expressions and certain Verb-Object patterns to influence article choice.

The whole analysis thus strengthens the idea that words are constructions and that language is captured by a complex network of constructions, i.e. the construct-i-con (Goldberg 2003: 219). Not only can the Construction Grammar framework describe linguistic phenomena; with a constructional approach, it is also possible to interpret and understand the relations happening among constructional elements in constructions that are not particularly frequent in a language. Finally, the use of a bottom-up approach effectively allowed the grounding for linguistic analysis based on corpus evidence.

7 Conclusions

7.1 Summary of results

The main goal of the current corpus-based study was to investigate article use in BrE, with a particular focus on bare noun phrases and article variability. In order to be able to obtain bare NPs with the potential to variable article use, the investigation made use of large parallel corpora. More specifically, by starting from a language that uses articles frequently and targeting NPs occurring with an article, parallel corpora allow the retrieval of the aligned bare NPs in another language. The language pair used in the current study was German and English. The data came from *Europarl* (*a Parallel Corpus for Statistical Machine Translation*), a collection of the proceedings of the parliamentary debates held at the European Parliament, and its improved version *CoStEP* (*Corrected & Structured Europarl Corpus*).

Originally, the proceedings collected in *Europarl* were not designed for linguistic purposes; therefore, the first part of the project focused on preliminary studies, which involved the evaluation of the corpus. The first analysis tested its reliability and faithfulness to the parliamentary transcriptions (see section 4.2). In order to do so, the transcripts of three debates were compared with the corresponding video recordings available online. Previous literature (see e.g. Mollin 2007) highlighted that transcripts differ from original speeches in terms of style (i.e. speakers' lexical and grammatical choices are modified and closer to written language); in other words, they are normally more conservative. The results confirmed the presence of discrepancies between the parliamentary transcripts and the politicians' original speeches. However, the differences were found at low frequency level, i.e. they corresponded to 4% for English and 1% for German. Due to the non-conventionalized character of some expressions used by the speakers, they are not yet accepted by standard grammars, and transcribers in both languages therefore prefer to move towards a more conservative style (with German being more conservative than English). Overall, the transcripts of *Europarl* are therefore considerably faithful to the speeches and can in turn be used for linguistic analysis.

A further obstacle of *Europarl* regarded the impossibility to distinguish between translated texts and original material. As claimed in previous studies (see e.g. Nisioi et al. 2016, Rabinovich et al. 2016, and Bernardini et al. 2016), this is

potentially problematic when investigating natural language use, as translations generally deviate from originals. Section 4.3 investigated whether *Europarl* translations are as trustworthy as original texts with respect to English article use. The analysis built on a second study (called Study A and discussed in more detail in Chapter 5), which compared article use between English and German with collective nouns. The difference between the two studies concerned the data retrieval process. While Study A used English originals, the study in question did not control the source language and therefore included three types of texts: English originals, English texts produced by non-native speakers, and translated material. Overall, the results of the analysis showed that the frequency of bare NPs in the data sample containing mixed texts was lower than the frequency attested in Study A (i.e. in which the speakers' mother tongue was controlled for). The findings thus supported the results of existing literature and strengthened the idea that translations are not fully reliable for the investigation of natural language use. Hence, in order to investigate language variation, the distinction between native and non-native speakers was necessary.

The study in Chapter 5 contrastively analysed article use in English and German with four collective nouns: *Parliament*, *Council*, *Committee*, and *people*. The main findings confirmed that German articles are used more extensively than in English. Moreover, the logistic regression analysis revealed that English article use is strongly influenced by the noun type and noun modification (see Table 5.2), i.e. article use with collective nouns is lexical-dependent and depends on pre- or post-modifying elements (e.g. *Ø Parliament* vs. *the European Parliament*). On the other hand, the syntactic function did not show any significant influence. Additionally, the study looked at the difference in article use between British and Irish speakers, as IrE has been reported to use the definite article more frequently. The results of the statistical analysis could not confirm this. Overall, therefore, the findings proved that using parallel corpora is an excellent method to explore similarities and differences between two (or more) languages and for the investigation of article use. More importantly, they demonstrated that German as a starting point was a suitable language for the current project.

Finally, Chapter 6 offered a bottom-up analysis and a closer look at English article use, by combining a data-driven method with a Construction Grammar perspective. The data-driven approach was particularly useful for the data selection, narrowing down the contexts in which an article could be potentially variable. The

retrieved bare data sample mainly contained uncountable nouns and plural count nouns. Furthermore, articles were usually omitted in NPs with generic reference. The results thus supported the general descriptions given by standard grammars in relation to article omission (see e.g. Quirk et al. 1985, Biber et al. 1999). Interestingly, as described by Rowlinson (1994: 87), the dataset also strengthened the notion that German abstract noun NPs are commonly preceded by an article, while in English they tend to occur as bare cases.

In order to allow for better focus, the constructional analysis narrowed down to the basic form of an NP (i.e. [[article] + [noun]]) and, therefore, did not include premodifying elements. In addition, it investigated an abstract nominal postmodified by an *of*-phrase (i.e. the *of*–CONSTRUCTION). The constructional account was strictly based on corpus evidence. The bottom-up approach thus allowed to build up more abstract generalizations (e.g. Langacker 1991, 2005), which were located at the micro-level and meso-level on the hierarchical abstraction continuum. The focus of the constructional representations was on the semantic properties within the single components and a construction as a whole. A first CxG modelling was based on the Determination Construction suggested by Fillmore (1988) and Fried and Östman (2004b). However, the proposed representations showed a few limitations and revealed some discrepancies that made the model incompatible with the tenets of Construction Grammar: the non-compositionality and non-predictability aspects (e.g. Goldberg 1995, 2006), as well as other potential phenomena that can occur within and among constructions, i.e. inheritance relations and coercion. A further relevant aspect regarded the representation of the empty determiner slot. The revised CxG model made use of the scale of definiteness discussed by Chesterman (1991), on which the zero article and the null article are located at opposite poles. The degree of definiteness is thus determined by the article: the zero and indefinite articles express [–definite], while the definite and null article express [+definite]. Based on this interpretation, both the zero and the null article therefore convey meaning, even though they are not overtly expressed. Hence, it was argued that the empty determiner slot in an NP is merely illusory, as it contains semantic properties and in turn contributes to the meaning of the construction. All articles, therefore, can be seen as constructions, and talking about article omission is not exactly correct. The constructional representations discussed in section 6.5 showed different types of NPs, determined by various interactions between two constructions, namely an article and a

noun (i.e. plural count noun, proper noun, and non-count/count abstract noun). The analysis showed that the constructional meaning varies depending on the interactions between the elements of a construction: articles, for instance, have meaning potential (e.g. specific, unique, familiar), but not all this meaning potential is realised in each construction. Taken together, therefore, the analysis effectively contributed to the notion that constructions are structured in a very complex constructional network (i.e. the construct-i-con), in which they are linked together via numerous nodes, exchanging different types of inheritance relations. Finally, the analysis proposed a constructional representation of the expression *the/Ø enlargement of the European Union*, in which article use shows variability. The meaning of the whole construction in both variants equals full specificity, but the internal semantic properties are differently structured. More specifically, the feature [+specific] is inherited from two different elements: either the definite article or the abstract noun that, in the bare form, turns into a proper noun, influenced by the external context. In order to account for this observed variability in some structures and abstracting from the presented construction frames, a constructional framework was presented. Adding the notion of allustructions to the construct-i-con proposes a plausible network element that explains why certain NPs only appear variable at the surface level, while others can be considered truly variable in the sense that they share a function, such as specificity.

To sum up, the CxG model proposed in the current study refined the existing models, with particular attention to the internal semantic relations and possible phenomena occurring within and among constructions (i.e. inheritance relations and coercion); furthermore, it suggested a constructional representation of bare NPs. This was possible by using empirical data (retrieved via a data-driven approach), from which the CxG model was created. A first look at evidence-based cases of article variability involving syntactic scope beyond the NP suggests that the proposed model can indeed be embedded into a construct-i-con that zooms out and represents more complex connections among constructions and abstractions beyond the analysed NP.

7.2 Limitations and further research

The initial lack of *Europarl* to differentiate original material from translated texts was later solved in *CoStEP*. One of *CoStEP*'s new functionalities was the possibility to access speakers' background information (e.g. nationality), which is generally the best

proxy available to distinguish original texts from translations. However, one source of weakness in this study that might have influenced the results was the unavoidable use of German translations for the retrieval of English original NPs. As already discussed, translations slightly differ from original texts. Consequently, German translations might be different from German original texts. Put differently, the data could thus be impacted due to the translation process. Since the study made use of parallel material, it is unfortunate that this limitation could not be totally fixed, as the focus was to analyse English texts produced by English native speakers (i.e. English originals). Therefore, it would be interesting to assess the effects of using German original texts and compare the results with the current study. It would be in turn possible to evaluate what types of data the retrieval process would produce in the equivalent parallel English translations. In order to make better assumptions on article use in English and explore more in the field of translation studies, future research would thus benefit from an additional investigation that used German originals as starting point. Hence, in general, the combination of the text type (i.e. transcriptions) and methodology (i.e. using German as a starting point) might have interfered with finding a higher number of novel bare cases in the corpus.

As the use of German as starting point mainly yielded abstract nouns, the study did not evaluate the investigation of (variable) article use with a larger variety of noun types (e.g. concrete nouns). Additionally, this did not allow a deeper analysis of the generic reference of all articles. As for instance described by Lyons (1999: 179-181), the generic reference is not exclusively expressed in a bare NP, but also with the definite and indefinite article (e.g. *the dog has four legs*, *a dog has four legs*, and *Ø dogs have four legs*). Furthermore, due to the scope of this project and the strictly evidence-based bottom-up analysis, the CxG model could only present a limited number of constructions and mainly focused on the micro-level and meso-level, i.e. it was not possible to thoroughly investigate higher levels of abstraction. In fact, the focus of the current study was between the boundaries within the noun phrase and did not take into account other factors that potentially influence article use, e.g. premodification, previously mentioned items (i.e. direct anaphoric reference), and elements external to the NP such as context or superimposed syntactic elements (e.g. preceding PP). The proposed constructional network did however try to address premodification based on the available data from Chapter 5 and thus laying out a path for future research to follow. Also, a handful of evidence based examples with scope

beyond the NP were discussed in section 6.6 in order to allow for a first glimpse of what might be taking place outside the NP and thus showing a path, how future research could deal with cases such as fixed expressions or VP influence. A larger and more varied sample will be helpful to investigate various and (more) complex noun phrases from a constructional point of view. In terms of directions for future research, further work needs to include the above-mentioned factors in order to have a more complete investigation and, therefore, to build a possible NP construction at the macro-level (i.e. the highest level of abstraction along the constructional hierarchy). Furthermore, this study raised an important question about the nature of article variability, namely whether it is possible to call it as such if two synonymous constructions inherit one feature from different positions. This aspect is definitely a stimulating point for further research, which will also be achievable with a larger and more stratified data sample.

Moreover, further investigation needs to be done to establish whether using a different language as a starting point (e.g. a Romance language) could give different results and yield other cases of English bare NP uses. As described by McIntosh (2002: 16) and Proudfoot and Cardo (2013: 13), for instance, Italian uses the definite article more consistently when referring to something generic (e.g. *Mi piace il gelato – I like Ø ice cream – Ich mag Ø Eiscrème* or *Le sigarette fanno male – Ø Cigarettes are bad for you – Ø Zigaretten sind schlecht für dich*). Furthermore, Italian uses articles less variably than English and German with weekdays, role predicates, and languages (McIntosh 2002: 170, Accademia della Crusca Online 2011, Zanichelli Online 2013), as shown in (1), (2), and (3), respectively.[106]

(1)
  (a) Ø Next Monday, Aung San Suu Kyi will be 55 years old [...]. (*CoStEP* 2000-06-15.xml)
  (b) Il prossimo lunedì Aung San Suu Kyi compirà 55 anni […].
  (c) Ø Nächsten Montag wird Aung San Suu Kyi 55 Jahre alt […].

(2)
  (a) As Ø President Obama said, it is not acceptable to put that many people's lives at risk. (*CoStEP* 2003-03-12.xml)
  (b) Come ha detto il presidente Obama, è inaccettabile che così tante persone rischino la vita.

[106] Note that the examples taken from *CoStEP* refer to English originals and Italian and German translations.

(c) Wie Ø Präsident Obama gesagt hat, ist es nicht akzeptabel, so viele Menschenleben in Gefahr zu bringen.

(3)

(a) You have to do that and the advantage of listening to Ø German is that all the verbs come at the end so that also helps! (*CoStEP* 2002-09-24.xml)
(b) Anche lei deve farlo, inoltre il vantaggio di ascoltare il tedesco è che tutti i verbi arrivano alla fine e anche questo è un aiuto!
(d) Es bleibt einem nichts weiter übrig, und Ø Deutsch hat zumindest den Vorteil, dass die Verben ganz am Schluss stehen.

Finally, McIntosh (2002: 17) mentions two more contexts in which Italian and English (but also German) tend to differ, namely with sports (e.g. *È fissata con il calcio – She's mad about Ø football – Sie ist verrückt nach Ø Fussball*)[107], and colours (e.g. *Il blu ti dona – Ø Blue suits you – Ø Blau steht dir gut*). To sum up, this brief overview proves that using another language in the methodology might yield diverse types of occurrences. Hence, it would be very interesting to repeat the study and compare the results with a different language as a starting point.

A final consideration needs to be made with respect to free variability. Other factors can influence the choice of the article, even if seemingly identical in meaning, such as regional variety (e.g. BrE vs. AmE), register (e.g. newspapers vs. academic writing), or context (e.g. headlines). These factors could not be investigated, as the used corpus is limited to parliamentary discourse and mostly BrE in a formal context.

To conclude, the present study made several noteworthy contributions. Firstly, this research has strengthened the idea that parallel corpora are a very useful and powerful tool for both linguistic analysis and the investigation of language variation. In particular, *CoStEP* provided an effective framework for the exploration of (variable) article use in English, showed great potential for the investigation of language variation, and can serve as a base for future studies in (Parallel) Corpus Linguistics. Secondly, the current study introduced and tested a new methodology for the retrieval of bare NPs that addresses the notorious difficulty of retrieving instances containing covert phenomena in corpus linguistics. Based on the empirical findings, the experimental method proved to be very valid. Thirdly, the study successfully corroborated the notion that corpora are beneficial for a bottom-up approach in the Construction Grammar framework (i.e. the Construction Grammar theory can be applied in the field of Corpus Linguistics). Finally, the present study adds to the

[107] An article, however, is not used with the verb *giocare*, as in *Gioco a Ø calcio* (*I play Ø football*).

growing body of research on article use/variation and Construction Grammar by challenging existing descriptive approaches and introducing more refined CxG models. Corpus-based studies as well as data-driven and constructional approaches should therefore remain in the focus for the investigation of linguistic phenomena, linguistic patterns, and linguistic variation.

References

Accademia della Crusca. 2011. *Omissione dell'articolo determinativo nella locuzione temporale* settimana prossima/scorsa. <http://www.accademiadellacrusca.it/it/lingua-italiana/consulenza-linguistica/domande-risposte/omissione-dellarticolo-determinativo-locuzio> (last accessed 5th July 2017)

Bailey, Charles-James N. 1987. Marginalia on Singulars and Plurals in English. *Arbeiten aus Anglistik und Amerikanistik*, 3-11.

Baroni, Marco, and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3): 259–274.

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis and Elena Tognini-Bonelli, eds., *Text and technology: In honour of John Sinclair*. Amsterdam: Benjamins, 233–250.

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target,* 7: 223–243.

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers, ed., *Terminology, LSP and translation*: *Studies in language engineering in honour of Juan C. Sager*. Amsterdam: Benjamins, 175–186.

Baker, Peter S. 2011. *Introduction to Old English*. *2nd Edition*. Oxford: Wiley-Blackwell.

Bell, Allan. 1985. One rule of news English: geographical, social and historical spread. *Te Reo*, 28: 95–117.

Bell, Allan. 1988. The British base and the American connection in New Zealand media English. *American Speech*, 63(4): 326–344.

Berezowski, Leszek. 2009. *The Myth of the Zero Article*. London: Continuum.

Bernardini, Silvia, Andiano Ferraresi, and Maja Miličević. 2016. From EPIC to EPTIC—Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, 28(1): 61–86.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Bies, Ann, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. *Bracketing guidelines for Treebank II style Penn Treebank project*. Department of Computer and Information Science: University of Pennsylvania.

Birner, Betty, and Gregory Ward. 1994. Uniqueness, familiarity, and the definite article in English. *Annual Meeting of the Berkeley Linguistics Society,* 20(1): 93–102.

Borin, Lars, ed., 2002. *Parallel Corpora, Parallel Worlds: Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22-23 April, 1999*. Amsterdam & New York: Rodopi.

Borin, Lars. 2002b. … and never the twain shall meet? *Language and Computers*, 43(1): 1–43.

Bucholtz, Mary. 2000. The politics of transcription. *Journal of Pragmatics*, 32(10): 1439–1465.

Burrow, John A., and Thorlac Turville-Petre. 2005. *A book of Middle English. 3rd Edition*. Oxford: Blackwell Publishing.

Callegaro, Elena, and Simon Clematide. 2017. *The validity of large data-driven and constructional approaches for the investigation of variable article use in English*. Paper presented at ICAME38. Charles University, Prague, 24–28 May 2017.

Callegaro, Elena, Simon Clematide, Marianne Hundt, and Sara Wick. 2019. Variable article use with acronyms and initialisms – a contrastive analysis of English, German and Italian. *Languages in contrast*, 19(1): 48–78.

Cappelle, Bert. 2006. Particle placement and the case for "allostructions". *Constructions online*, 1(7): 1–28.

Chesterman, Andrew. 1991. *On Definiteness: A Study with Special Reference to English and Finnish*. Cambridge: Cambridge University Press.

Church, Kenneth W., and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational linguistics*, 19(1): 1–24.

Christophersen, Paul. 1939. *The Articles: a Study of their Theory and Use in English*. Copenhagen: Munksgaard.

Clematide, Simon, Johannes Graën, and Martin Volk. 2016. Multilingwis – A Multilingual Search Tool for Multi-Word Units in Multiparallel Corpora. In Gloria Corpas Pastor, ed., *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives/Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües*. Geneva: Tradulex, 447–455.

Collins Online. 2017. *The Collins English Dictionary*. <http://www.collinsdictionary.com/dictionary/english/article> (last accessed 19th July 2017)
<http://www.collinsdictionary.com/dictionary/english/collective-noun#collective-noun_1> (last accessed 4th May 2017)

Codrea-Rado, Anna. 21 May 2014. European parliament has 24 official languages, but MEPs prefer English. *The Guardian*. <http://www.theguardian.com/education/datablog/2014/may/21/european-parliament-english-language-official-debates-data> (last accessed 18th August 2015)

Corrigan, Karen P. 2010. *Irish English: Northern Ireland. Vol. 1*. Edinburgh: Edinburgh University Press.

Coulthard, Malcolm. 1996. The official version: audience manipulation in police records of interviews with suspects. In Carmen-Rosa Caldas-Coulthard and Malcolm Coulthard, eds., *Texts and practices: readings in critical discourse analysis*. London: Routledge, 166-178.

Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago: University of Chicago Press.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Croft, William. 2005. Logical and typological arguments for Radical Construction Grammar. In Jan-Ola Östman and Mirjam Fried, eds., *Construction Grammars: Cognitive grounding and theoretical extensions*. Amsterdam: Benjamins, 273–314.

Cruse, D. Alan. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

Cruse, D. Alan. 2000. *Meaning in language: an introduction to semantics and pragmatics*. Oxford: Oxford University Press.

Crystal, David. 2008. *A dictionary of linguistics and phonetics. 6th Edition*. Oxford: Blackwell Publishing.

Curme, George O. 1970. *A grammar of the German language*. *2nd Revised Edition*. New York: Ungar.

Depraetere, Ilse. 2003. On verbal concord with collective nouns in British English. *English Language and Linguistics*, 7(1): 85–127.

Dryer, Matthew S., and Martin Haspelmath, eds., 2013. *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, last accessed 11th June 2015)

Dryer, Matthew S. 2013. Definite Articles. In Matthew S. Dryer and Martin Haspelmath, eds., *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/37, last accessed 11th June 2015)

Dryer, Matthew S. 2013. Indefinite Articles. In Matthew S. Dryer and Martin Haspelmath, eds., *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/38, last accessed 11th June 2015)

Dudenredaktion, ed., 2005. *Duden – die Grammatik: unentbehrlich für richtiges Deutsch, 7th Edition*. *Vol. 4*. Mannheim: Dudenverlag.

Duden Online. 2017. *Die Verteilung der Artikel (Genusangabe) im Rechtschreibduden.*
<http://www.duden.de/sprachwissen/sprachratgeber/die-verteilung-der-artikel-genusangabe-im-rechtschreibduden> (last accessed 28th June 2017)

European Parliament
<http://www.europarl.europa.eu/portal/en> (last accessed 26th November 2016)

European Parliament Plenary – *Debates and Videos*
<http://www.europarl.europa.eu/plenary/en/debates-video.html> (last accessed 5th March 2017)

European Parliament – *Never lost in translation*
<http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+IM-PRESS+20071017FCS11816+0+DOC+XML+V0//EN> (last accessed 18th August 2016)

European Parliament – *European Commission, Supporting language learning and linguistic diversity*
<http://ec.europa.eu/languages/policy/linguistic-diversity/official-languages-eu_en.htm> (last accessed 18th August 2016)

European Parliament – *List of speakers' names*
<http://www.europarl.europa.eu/ep-live/en/plenary/search-by-organ?legislature=-1&country=GB&group=&type_organ=all> (last accessed 25th February 2017)

European Parliament Proceedings Parallel Corpus 1996-2011
<http://www.statmt.org/europarl/> (last accessed 15th March 2017)

Filppula, Markku. 2008. Irish English: morphology and syntax. *Varieties of English*, 1: 328–359.

Fillmore, Charles J. 1968. The case for case. In Emma Bach and Robert T. Harms, eds., *Universal in Linguistic Theory*. New York: Holt, Rinehart and Winston, 1–88.

Fillmore, Charles J. 1982. Frame Semantics. In Linguistic Society of Korea, ed., *Linguistics in the Morning Calm*. Seoul: Hanshin, 111–138.

Fillmore, Charles J. 1988. The mechanisms of "construction grammar". In Shelley Axmaker, Annie Jaisser and Helen Singmaster, eds., *Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society, 35–55.

Frege, Gottlob. 1967. Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought. In Jean van Heijenoort, ed., *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Cambridge: Harvard University Press, 1–82.

Fried, Mirjam, and Jan-Ola Östman 2004a. *Construction Grammar in a cross-language perspective*. Amsterdam: Benjamins.

Fried, Mirjam, and Jan-Ola Östman. 2004b. Construction Grammar: a thumbnail sketch. In Mirjam Fried and Jan-Ola Östman, eds., *Construction Grammar in a cross-language perspective*. Amsterdam: Benjamins, 11–86.

Fries, Udo. 1988. The crew have abandoned the ship: Concord with collective nouns revisited. *Arbeiten aus Anglistik und Amerikanistik*, 99–104.

Gao, Qin, and Stephan Vogel. 2008. Parallel implementations of word alignment tool. *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, 49–57.

Ginzburg, Jonathan, and Ivan A. Sag. 2000. *Interrogative investigations*. Stanford: CSLI publications.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2002. Surface generalizations: An alternative to alternations. *Cognitive Linguistics*, 13(4): 327–356.

Goldberg, Adele E. 2003. Constructions: a new theoretical approach to language. *Trends in cognitive sciences*, 7(5): 219–224.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press on Demand.

Goldberg, Adele E. 2013. Constructionist approaches. In Thomas Hoffmann and Graeme Trousdale, eds., *The Oxford handbook of construction grammar*. Oxford: Oxford University Press, 15–31.

Graën, Johannes, Dolores Batinic, and Martin Volk. 2014. *Cleaning the Europarl Corpus for Linguistic Applications*. Paper presented at Konvens 2014, University of Hildesheim, Hildesheim, 8–10 October 2014.

Graën, Johannes. 2017. Identifying phrasemes via interlingual association measures - A data-driven approach on dependency-parsed and word-aligned parallel corpora. In Christine Konecny, Erica Autelli, Andrea Abel, and Lorenzo Zanasi, eds., *Lexemkombinationen und typisierte Rede im mehrsprachigen Kontext*. Tübingen: Stauffenburg Verlag, in print.

Graën, Johannes. 2018. *Exploiting alignment in multiparallel corpora for applications in linguistics and language learning.* University of Zurich, Faculty of Arts.

Green, Judith, Maria Franquiz, and Carol Dixon. 1997. The myth of the objective transcript: Transcribing as a situated act. *Tesol Quartertly*, 31(1): 172–176.

Greenbaum, Sidney, ed., 1996. *Comparing English worldwide: The international corpus of English*. Oxford: Clarendon Press.

Halliday, Michael A. 1989. *Spoken and written language*. Oxford: Oxford University Press.

Harley, Heidi. 2004. Why is it the CIA but not *the NASA? Acronyms, initialisms, and definite descriptions. *American Speech*, 79(4): 368–399.
Hentschel, Elke, ed., 2010. *Deutsche Grammatik*. Berlin: Walter de Gruyter.
Hickey, Raymond. 2007. *Irish English: History and present-day forms*. Cambridge: Cambridge University Press.
Hilpert, Martin. 2014. *Construction grammar and its application to English*. Edinburgh: Edinburgh University Press.
Hoffmann, Thomas, and Graeme Trousdale, eds., 2013. *The Oxford handbook of construction grammar*. Oxford: Oxford University Press.
Hollmann, Willem B., and Anna Siewierska. 2011. The status of frequency, schemas, and identity in Cognitive Sociolinguistics: A case study on definite article reduction. *Cognitive Linguistics*, 22(1): 25-54.
Huddleston, Rodney, and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English language*. Cambridge: Cambridge University Press.
Hundt, Marianne. 2016. *Who is the/a/ø professor at your university?* A construction-grammar view on changing article use with single role predicates in American English. In María José López-Couso, Bélen Méndez-Naya, Paloma Núñez-Pertejo and Ignacio M. Palacios-Martínez, eds., *Corpus Linguistics on the Move: Exploring and Understanding English Through Corpora*. Amsterdam & New York: Brill/Rodopi, 227–258.
Hundt, Marianne. 2018. Variable article usage with institutional nouns – an 'oddment' of English? In Alex Ho-Cheong Leung and Wim van der Wurff, eds., *The Noun Phrase in English: Past and Present*. Amsterdam: Benjamins, 113–142.
Ilisei, Iustina, Diana Inkpen, Gloria Corpas Pastor, and Rusland Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, ed., *Proceedings of CICLing-2010: Computational Linguistics and Intelligent Text Processing*. New York: Springer, 503–551.
Jackendoff, Ray. 2003. *Foundations of language: brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
Jackendoff, Ray. 2013. Constructions in the parallel architecture. In Thomas Hoffmann and Graeme Trousdale, eds., *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 70–92.
Jespersen, Otto. 1949. *A Modern English Grammar on Historical Principles. Part VII*. Copenhagen: Munksgaard.
Johansson, Stig, and Knut Hofland. 1989. *Frequency Analysis of English Vocabulary and Grammar: Based on the LOB Corpus. Vol. 2*. Oxford: Oxford University Press.
Johansson, Stig. 2002. Towards a multilingual corpus for contrastive analysis and translation studies. *Language and Computers*, 43(1): 47–59.
Johansson, Stig. 2007. Seeing through multilingual corpora. *Language and Computers*, 62(1): 51–71.
Jucker, Adreas H. 1992. *Social Stylistics. Syntactic Variation in British Newspapers*. Berlin: Mouton de Gruyter.
Kallen, Jeffrey F. 2013. *Irish English Volume 2: The Republic of Ireland*. Berlin: Walter de Gruyter.
Kay, Paul, and Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: the What's X doing Y? construction. *Language*, 75(1): 1–33.
Kellerman, Eric, and Michael Sharwood Smith. 1986. *Crosslinguistic influence in second language acquisition*. New York: Pergamon Institute of English.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London: Longman.
Koch, Peter, and Wulf Oesterreicher. 1985. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschicht. Langage de la proximité – langage de la distance. L'oralité et la scripturalité entre la théorie linguistique et l'histoire de la langue. *Romanistisches Jahrbuch*, 36: 15–43.
Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5: 79–86.
König, Ekkehard, and Volker Gast. 2009. *Understanding English-German contrasts*. Berlin: Schmidt Erich Verlag.
Kruisinga, Etsko, and Pieter A. Erades. 1960. *An English grammar: Accidence and syntax*. *Vol. 1*. Groningen: Noordhoff.
Kučera, Henry, and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Providence: Brown University Press.
Kurokawa, David, Cyril Goutte, and Pierre Isabelle. 2009. *Automatic detection of translated text and its impact on machine translation*. Paper presented at the 12th Machine Translation Summit.
<http://www.mt-archive.info/MTS-2009-Kurokawa.pdf> (last accessed 5th April 2020)
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language*, 45(4): 715–762.
Labov, William. 1994. *Principles of linguistic change. Vol. 1. Internal factors*. Oxford: Blackwell Publishing.
Labov, William. 2001. *Principles of linguistic change. Vol. 2. Social factors*. Oxford: Blackwell Publishing.
Lakoff, George. 1987a. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.
Lakoff, George. 1987b. Cognitive models and prototype theory. In Ulric Neisser, ed., *Concepts and Conceptual Development*. Cambridge: Cambridge University Press, 391–421.
Langacker, Ronald W. 1987. *Foundations of cognitive grammar: Theoretical prerequisites. Vol. 1*. Stanford: Stanford University Press.
Langacker, Ronald W. 1991. *Foundations of cognitive grammar: Descriptive application. Vol. 2*. Stanford: Stanford University Press.
Langacker, Ronald W. 1999. *Grammar and conceptualization*. *Vol. 14*. Berlin: Walter de Gruyter.
Langacker, Ronald W. 2005. Construction Grammars: cognitive, radical, and less so. In Francisco J. Ruiz de Mendoza Ibáñez and M. Sandra Peña Cervel, eds., *Cognitive linguistics: Internal dynamics and interdisciplinary interaction*, Berlin: Mouton de Gruyter, 101–159.
Laviosa-Braithwaite, Sara. 1996. Comparable corpora: Towards a corpuslinguistic methodology for the empirical study of translation. In Marcel Thelen and Barbara Lewandowska Tomaszczyk, eds., *Translation and meaning. Part 3*. Maastricht: Rijkshogeschool, 153–163.
Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. *Directions in corpus linguistics: proceedings of Nobel symposium 82, Stockholm, 4–8 August*. Berlin: Mouton de Gruyter, 105–122.
Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith. 2009. *Change in contemporary English: A grammatical study*. Cambridge: Cambridge

University Press.
Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4): 799–825.
Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2013. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4): 999–1023.
Levin, Magnus. 2006. Collective nouns and language change. *English Language and Linguistics*, 10(2): 321–343.
Liang, Percy, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In Robert C. Moore (Conference Chair), *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 104–111.
Lyons, Christopher. 1991. Reference and articles. In Gerhard Leitner, ed., *English traditional grammars: an international perspective*. Amsterdam: Benjamins, 309–328.
Mauranen, Anna. 1999. Will 'translationese' ruin a contrastive study? *Languages in contrast*, 2(2): 161–185.
McIntosh, Collin. 2002. *The Oxford Italian grammar and verbs*. Oxford: Oxford University Press.
Michaelis, Laura A., and Knud Lambrecht. 1996. Toward a construction-based theory of language function: The case of nominal extraposition. *Language*, 72(2): 215–247.
Michaelis, Laura A. 2004. Type shifting in Construction Grammar: An integrated approach to aspectual coercion. *Cognitive Linguistics*, 15(1): 1–67.
Michaelis, Laura A. 2013. Sign-based construction grammar. In Thomas Hoffmann and Graeme Trousdale, eds., *The Oxford Handbook of Construction Grammar*, Oxford: Oxford University Press, 133–152.
Mollin, Sandra. 2007. The Hansard hazard: gauging the accuracy of British parliamentary transcripts. *Corpora,* 2(2): 187–210.
Montague, Richard. 1974. The Proper Treatment of Quantifiers in Ordinary English. In Richmond H. Thomason, ed., *Formal Philosophy*. New Haven: Yale University Press, 247–270.
Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli, and Mariachiara Russo. 2005. Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta: Translators' Journal*, 50(4): 114–129.
Nisioi, Sergiu, Ella Rabinovich, Liviu P. Dinu, and Shuly Wintner. 2016. A corpus of native, non-native and translated texts. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis, eds., *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, Paris: European Language Resources Association (ELRA), 4197–4201.
Nivre, Joakim, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In Nicoletta Calzolari (Conference Chair), Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk and Daniel Tapias, eds., *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Paris: European Language Resources Association

(ELRA), 2216–2219.
O'Connell, Daniel C., and Sabine Kowal. 1999. Transcription and the issue of standardization. *Journal of Psycholinguistic research*, 28(2): 103–120.
Odlin, Terence. 1989. *Language transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.
Osherson, Daniel N., and Edward E. Smith. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9(1): 35–58.
Øverås, Linn. 1998. In search of the third code: An investigation of norms in literary translation. *Meta: Translators' Journal*, 43(4): 557–570.
Partington, Alan. 1998. *Patterns and meanings: Using corpora for English language research and teaching. Vol. 2*. Amsterdam: Benjamins.
Perek, Florent. 2012. Alternation-based generalizations are stored in the mental grammar: Evidence from a sorting task experiment. *Cognitive Linguistics*, 23(3): 601–635.
Platt, John T. 1974. Alphabet Soups or a Mess of Pottage? *Foundations of Language*, 11(2): 295–297.
Platt, John T., Heidi Weber, and Mian Lian Ho. 1984. *The New Englishes*. London: Routledge & Kegan Paul.
Poutsma, Hendrik. 1904. *A grammar of late modern English, for the use of continental, especially Dutch, students. Part II The parts of speech*. Groningen: Noordhoff.
Proudfoot, Anna, and Francesco Cardo. 2013. *Modern Italian grammar: a practical guide. 3rd Edition*. London & New York: Routledge.
Puurtinen, Tiina. 2003. Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing*, 18(4): 389–406.
Pym, Anthony, François Grin, Claudio Sfreddo, and Andy LJ Chan. 2013. *The status of the translation profession in the European Union*. London: Anthem Press.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
Rabinovich, Ella, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the Similarities Between Native, Non-native and Translated Texts. In Katrin Erk and Noah A. Smith, eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg: Association for Computational Linguistics, 1870–1881.
Rosch, Eleanor H. 1973. Natural categories. *Cognitive psychology*, 4(3): 328–350.
Rosch, Eleanor H., Carol Simpson, and R. Scott Miller. 1976. Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, 2(4): 491–502.
Rosch, Eleanor H., and Barbara Bloom Lloyd, eds., 1978. *Cognition and categorization. Vol. 1*. New York: Lawrence Erlbaum Associates.
Ross, John R. 1972. Alphabet soups and name-calling. *Foundations of Language*, 9(1): 113.
Rowlinson, William. 1994. *German grammar*. Oxford: Oxford University Press.
Rydén, Mats. 1975. Noun-name collocations in British English newspaper language. *Studia Neophilologica*, 47: 14–39.
Sag, Ivan A. 1997. English relative clause constructions. *Journal of linguistics*, 33(2): 431–483.
Sand, Andrea. 2004. Shared morpho-syntactic features in contact varieties of English:

Article use. *World Englishes*, 23(2): 281–298.
Schmid, Helmut. 1994. *Probabilistic Part-Of-Speech Tagging Using Decision Trees*. Paper presented at the International Conference on New Methods in Language Processing.
<https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf> (last accessed 5th April 2020)
Sharwood Smith, Michael, and Eric Kellerman. 1986. Crosslinguistic influence in second language acquisition: An introduction. In Eric Kellerman and Micheal Sharwood Smith, eds., *Crosslinguistic influence in second language acquisition*. Oxford: Pergamon Press, 1–9.
Siemund, Peter. 2013. *Varieties of English: a typological approach*. Cambridge: Cambridge University Press.
Sinclair, John M. 1992. The automatic analysis of corpora. In Jan Svartvik, ed., *Directions in Corpus Linguistics: proceedings of Nobel symposium 82, Stockholm, 4–8 August*. Berlin: Mouton de Gruyter, 379–397.
Slembrouck, Stef. 1992. The parliamentary Hansard 'verbatim' report: the written construction of spoken discourse. *Language and Literature*, 2: 101–119.
Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.
Taylor, John R. 2002. *Cognitive grammar*. Oxford: Oxford University Press.
Taylor, John R. 2003. *Linguistic categorization*. *3rd Edition*. Oxford: Oxford University Press.
Tannen, Deborah. 1982. *Spoken and written language: Exploring orality and literacy*. New York: Ablex.
Teich, Elke. 2003. *Cross-Linguistic Variation in System and Text*. Berlin: Mouton de Gruyter.
Teubert, Wolfgang. 1996. Comparable or parallel corpora? *International Journal of Lexicography*, 9(3): 238–264.
Tirkkonen-Condit, Sonja. 2002. Translationese – a myth or an empirical fact? A study into the linguistic identifiability of translated language. *Target*, 14(2): 207–220.
Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins.
Traugott, Elizabeth C. 2008. Grammaticalization, constructions and the incremental development of language: Suggestions from the development of degree modifiers in English. In Regine Eckhardt, Gerhard Jäger and Tonjes Veenstra, eds., *Variation, Selection, Development – Probing the Evolutionary Model of Language Change*. Berlin: Mouton de Gruyter, 219–250.
Trousdale, Graeme. 2008. A constructional approach to lexicalization processes in the history of English: Evidence from possessive constructions. *Word Structure*, 1: 156–177.
Trousdale, Graeme. 2010. *An introduction to English sociolinguistics*. Edinburgh: Edinburgh University Press.
Tse, Grace Y.W. 2001. The grammatical factors influencing the choice between the use and omission of the definite article preceding multi-word organization names: a statistical analysis. *Journal of Quantitative Linguistics*, 8(1): 13–32.
Tse, Grace Y.W. 2003. Validating the logistic model of article usage preceding multi-word organization names with the aid of computer corpora. *Literary and Linguistic Computing,* 18(3): 287–313.
Tse, Grace Y.W. 2004. A grammatical study of personal names in present-day English: with special reference to the usage of the definite article. *English studies*, 85(3): 241–259.

Twitto-Shmuel, Naama, Noam Ordan, and Shuly Wintner. 2015. Statistical machine translation with automatic identification of translationese. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva and Pavel Pecina, eds., *Proceedings of the 10th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics, 47–57.

Van Halteren, Hans. 2008. Source language markers in EUROPARL translations. In Donia Scott and Hans Uszkoreit, eds., *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester: Coling 2008 Organizing Committee, 937–944.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov and Nikolai Nikolov, eds., *Proceedings of International Conference Recent Advances in Natural Language Processing (RANLP 2005)*. Shoumen: INCOMA, 590–596.

Volansky, Vered, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1): 98–118.

Walker, Graffam. 1990. *Language at work in the law: the customs, conventions, and appellate consequences of court reporting*. In Levi Walker and Graffam Walker, eds., L*anguage in the judicial process*. New York: Plenum, 203-244.

Willems, Dominique, Bart Defrancq, Timothy Colleman, and Dirk Noël. 2004. *Contrastive analysis in language: identifying linguistic units of comparison*. London: Palgrave Macmillan.

Yoo, Isaiah W. 2007. Definite article usage before last/next time in spoken and written American English. *International Journal of Corpus Linguistics*, 12(1): 83–105.

Yoo, Isaiah W. 2009. The English definite article: What ESL/EFL grammars say and what corpus findings show. *Journal of English for Academic Purposes*, 8(4): 267–278.

Zanettin, Federico. 1998. Bilingual comparable corpora and the training of translators. *Meta: Translators' Journal*, 43(4): 616–630.

Zanichelli Online. 2013. *Uso dell'articolo determinativo.* <http://aulalingue.scuola.zanichelli.it/benvenuti/2009/11/26/uso-dellarticolo-determinativo/> (last accessed 5th July 2017)

Appendices

Appendix A

Frequency list of the syntactic pattern distribution over *CoStEP*.

Retrieval pattern	Frequency
prepositional phrase (noun)	63,209
prepositional phrase (verb)	18,147
object phrase	17,895
subject phrase	14,689
coordinated phrase	12,157
non-finite phrase	8,010
passive subject phrase	4,784
predicative clause	1,692

Appendix B

Complete list of singular and abstract nouns included in the data set of bare NPs.

Abhorrence, abortion, abuse, acceptability, access, accessibility, accession, accountability, acidification, action, activity, adaptability, adherence, admissibility, advice, agreement, agriculture, aid, appeasement, appreciation, area, arrest, assurance, attendance, attention, authorization, awareness, bailout, bankruptcy, beer, board, breakdown, bureaucracy, business, capital, capitalism, care, caution, change, chaptalisation, choice, clarification, coherence, cohesion, comitology, commerce, commitment, comparability, compensation, competence, competition, compliance, compromise, concentration, conciliation, confidence, consensus, consolidation, consultation, control, cooperation, coordination, creation, credibility, crime, crisis, cultivation, culture, danger, date, death, decision-making, demand, democracy, desire, development, dignity, disability, disagreement, disapproval, disarmament, discharge, discipline, discrimination, disequilibrium, dissemination, diversity, dream, education, effect, efficiency, effort, eligibility, emphasis, encouragement, enforcement, enlargement, enterprise, equality, eradication, ethos, evidence, evil, exchange, exclusion, exploitation, exploration, extrapolation, farming, fatigue, fishing, follow-through, food, fortitude, freedom, frustration, funding, fusion, gambling, governance, gravitas, growth, harmonisation, health, history, hypocrisy, immigration, immunity, implementation, inability, income, industry, inflation, information, innovation, instability, integration, interaction, intolerance, investment, jingoism, justice, knowledge, lack, largesse, laundering, law, leadership, learning, legislation, legitimacy, level, life, linkage, logistics, loss, management, marriage, membership, mention, migration, mobility, modernisation, modulation, monitoring, movement, nature, negotiation, news, nonsense, occupation, openness, operation, opinion, opportunity, orientation, origin, oversight, paperwork, participation, peace, peace-building, persecution, place, plant, pleasure, policy, politics, pollution, power, practice, preservation, pressure, prevention, privatisation, privatization, procurement, production, progress, promotion, propaganda, property, proportionality, prosperity, prostitution, protection, racism, ratification, re-distribution,

reality, reassurance, recognition, reconciliation, reconstruction, referral, reform, regeneration, regression, regulation, relevance, reliance, remuneration, research, resolution, respect, responsibility, revolution, risk, safety, sanitation, satisfaction, scrutiny, security, sense, sharing, sight, simplification, smoking, society, solidarity, space, spending, sponsorship, spotlight, stability, status, sterling, strategy, success, supply, support, surrender, surveillance, suspension, sustainability, system, taxation, television, tenacity, territory, terrorism, test, tolerance, traceability, track, trade, trading, trafficking, training, transparency, transportation, travel, turbulence, unemployment, universality, use, value, vanity, violence, warming, wastage welfare, will, worry.

Appendix C

Complete frequency list of the nouns included in the follow-up analysis.

	Lemma	**Frequency**
1	policy	1780
2	industry	1518
3	development	1468
4	legislation	1423
5	action	1227
6	information	1196
7	law	1192
8	system	1184
9	support	1159
10	agreement	1122
11	level	1093
12	resolution	1088
13	access	1056
14	cooperation	1011
15	protection	968
16	safety	962
17	change	914
18	democracy	913
19	use	888
20	progress	887
21	security	878
22	crisis	859
23	reform	800
24	aid	776
25	regulation	772
26	enlargement	761
27	health	753
28	opportunity	711
29	research	702
30	trade	674
31	peace	650
32	responsibility	649
33	funding	643
34	area	634
35	society	631
36	implementation	595
37	growth	580
38	control	544
39	life	538
40	education	530
41	freedom	526
42	business	515
43	terrorism	496
44	violence	495
45	food	494
46	transparency	490
47	competition	480
48	agriculture	479
49	investment	478
50	confidence	478
51	respect	467
52	production	467
53	power	467
54	management	464
55	strategy	461
56	pressure	441
57	practice	441
58	value	412
59	evidence	412
60	crime	404
61	movement	403

62	success	392
63	risk	386
64	compromise	385
65	commitment	384
66	justice	380
67	creation	368
68	effect	363
69	stability	356
70	opinion	356
71	reality	355
72	training	354
73	unemployment	348
74	integration	341
75	efficiency	341
76	discrimination	327
77	attention	316
78	recognition	282
79	nature	266
80	danger	266
81	history	262
82	innovation	260
83	membership	259
84	discharge	256
85	sense	252
86	consultation	251
87	coordination	246
88	accountability	240
89	culture	238
90	consensus	233
91	fishing	230
92	spending	228
93	status	227
94	lack	227
95	choice	222
96	supply	221
97	abuse	215
98	loss	212
99	equality	210
100	solidarity	205
101	death	203
102	pollution	201
103	compensation	201
104	will	199
105	welfare	198
106	emphasis	198

107	prosperity	197
108	place	196
109	governance	195
110	cohesion	191
111	effort	187
112	demand	187
113	taxation	186
114	operation	186
115	participation	182
116	conciliation	182
117	accession	182
118	capital	179
119	racism	177
120	knowledge	172
121	activity	172
122	advice	170
123	care	169
124	promotion	168
125	bureaucracy	168
126	farming	163
127	harmonisation	161
128	monitoring	159
129	leadership	159
130	competence	157
131	diversity	155
132	date	153
133	exclusion	152
134	immigration	150
135	trafficking	149
136	origin	149
137	sustainability	147
138	enforcement	142
139	prevention	137
140	scrutiny	133
141	openness	131
142	news	126
143	ratification	124
144	compliance	124
145	exploitation	122
146	exchange	120
147	mobility	114
148	migration	108
149	income	107
150	pleasure	104
151	board	102

152	plant	99
153	reconciliation	97
154	disability	96
155	awareness	96
156	legitimacy	95
157	decision-making	94
158	credibility	94
159	dignity	92
160	smoking	89
161	coherence	89
162	reconstruction	88
163	clarification	86
164	test	84
165	television	84
166	negotiation	84
167	immunity	83
168	warming	82
169	procurement	82
170	space	80
171	arrest	79
172	territory	77
173	concentration	74
174	desire	73
175	instability	72
176	commerce	72
177	enterprise	71
178	simplification	70
179	assurance	69
180	inflation	67
181	traceability	66
182	property	64
183	travel	63
184	revolution	63
185	sight	62
186	learning	61
187	suspension	60
188	persecution	60
189	trading	59
190	hypocrisy	58
191	abortion	57
192	tolerance	56
193	modernisation	55
194	surveillance	52
195	caution	52
196	discipline	51

197	proportionality	49
198	nonsense	49
199	encouragement	48
200	transportation	47
201	gambling	47
202	eradication	47
203	orientation	46
204	disarmament	45
205	oversight	43
206	relevance	42
207	occupation	41
208	sharing	39
209	preservation	38
210	frustration	38
211	consolidation	37
212	laundering	36
213	breakdown	35
214	capitalism	34
215	track	31
216	prostitution	31
217	propaganda	31
218	modulation	31
219	bankruptcy	31
220	paperwork	29
221	attendance	29
222	satisfaction	28
223	marriage	27
224	dissemination	27
225	disagreement	27
226	regeneration	26
227	intolerance	26
228	evil	25
229	accessibility	25
230	reassurance	24
231	interaction	24
232	fusion	24
233	referral	23
234	eligibility	22
235	dream	22
236	appreciation	20
237	ethos	18
238	inability	17
239	remuneration	16
240	worry	15
241	privatization	15

242	mention	15
243	universality	14
244	linkage	14
245	spotlight	13
246	exploration	13
247	acceptability	13
248	surrender	12
249	turbulence	11
250	reliance	11
251	cultivation	11
252	adherence	11
253	wastage	10
254	sponsorship	10
255	sanitation	10
256	beer	10
257	sterling	9
258	politics	8
259	fatigue	8
260	extrapolation	8
261	bailout	8
262	appeasement	8
263	adaptability	8
264	tenacity	5
265	regression	5
266	jingoism	5
267	disapproval	5
268	comparability	5
269	admissibility	5
270	largesse	4
271	fortitude	4
272	authorization	4
273	abhorrence	3
274	vanity	2
275	disequilibrium	1
276	acidification	1